

**PhD Thesis**

**Predictive validity of the examination for the  
Membership of the Royal Colleges of Physicians  
of the United Kingdom**

**Katarzyna Ludka-Stempień**

**Medical School  
University College London**

**Supervisors:  
Professor Chris McManus and Dr Katherine Woolf**

**This thesis has been submitted in partial fulfilment of the requirements for  
the degree of Doctor of Philosophy**

**September 10<sup>th</sup>, 2014  
(amended April 21<sup>st</sup>, 2015)**



## Declaration

I, Katarzyna Ludka-Stempień, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature: \_\_\_\_\_

A handwritten signature in blue ink, appearing to be 'K. Ludka-Stempień', written over a horizontal line.

Date: 21/04/2015



## Abstract

*The constant public demand for high-quality medical services drives an associated demand for professional doctors, and requires them to take high-stakes exams. MRCP(UK) is a major examination for physicians in the UK, which aims to assess their knowledge, skills and appropriate attitudes – all key aspects of being a medical professional. Although the existing literature provides extensive evidence supporting the quality of MRCP(UK), this research aimed to add to the existing body of knowledge by investigating the predictive validity of MRCP(UK), i.e. examining whether it truly selects candidates who possess the above-mentioned qualities. This research therefore investigated the relationships between MRCP(UK) scores and results of seventeen knowledge exams and two clinical skills assessments (including specialty exams and MRCGP), training performance assessment outcomes (ARCP), and cases of licence limitations and erasures. Operating with the hypothesis that MRCP(UK) would predict all of the above-mentioned criteria, a retrospective longitudinal approach was assumed. The main sample contained records of 50,311 MRCP(UK) candidates attempting MRCP(UK) between May 2003 and January 2011; however, the analyses were performed on smaller samples, from 8 to 33,359 cases, depending on the size of the criterial dataset. The results of univariate and multivariate analyses supported the hypothesis. MRCP(UK) scores were indeed predictive of results of all knowledge exams and clinical assessments (meta-analysed average effects:  $r=0.69$  for Part I,  $r=0.70$  for Part II,  $0.48$  for PACES), and of performance in specialty training and issues with the licence to practice (on average:  $r=0.24$  for Part I, and  $r=0.22$  for Part II and PACES). The magnitudes of these validity coefficients were consistent with the theoretical notions of psychometrics and concurred with the findings of published studies. In view of the evidence it was concluded that MRCP(UK) is a valid exam. The limitations of this study, directions for future research, and general implications were discussed.*



## Acknowledgements

*Firstly, I would like to thank both of my superb supervisors Professor Chris McManus and Dr Katherine Woolf, who not only provided me with academic guidance, but were a constant and inexhaustible source of inspiration and support. Kath and Chris, I will miss your honest and wise advice, the kindness and welcoming approach you always presented, and the readiness to give me all the help I needed to be successful with this project. Additionally, I wish to thank Chris for all the intellectual feasts we shared; although they were often unexpectedly off-topic, they were always enlightening and motivating at the same time.*

*I would also like to thank my husband, friends, and academic and work colleagues who consistently supported me throughout this process, and who at the moments of self-doubt believed in me more than I did believe in myself. It is with their explicit or implicit support this research was accomplished. In particular, I want to thank Sam, Nilay, and Lynne for their advice in my struggles with the complexities of the English language, Alison, Henry, and Ola for their critical approach to my work, and Mel whose kindness got me through the most difficult adjustment period and then accompanied me throughout the entire PhD experience.*

*Lastly, I wish to express my extensive and foremost gratitude to the Royal Colleges of Physicians for funding this PhD project in cooperation with UCL IMPACT, and for their support in acquiring the necessary data. The project would not have been possible without the explicit involvement of Professor Jane Dacre, who supported it from the beginning, and Liliana Chis, whose comprehensive knowledge helped me understand the MRCP(UK) as a process. Furthermore, this PhD would not have been completed without the support of the General Medical Council, the Joint Royal Colleges of Physicians Training Board, the Royal College of Radiologists, the Royal College of General Practitioners, and the British Cardiovascular Society, who all kindly agreed to grant me access to their data.*

*This PhD project was a great opportunity for me to apply my psychometric knowledge in service of public interest, which was an extraordinary experience; one I will never forget. Thank you all for that.*

*London, September 10<sup>th</sup>, 2014*



## Table of Contents

<b>Declaration .....</b>	<b>3</b>
<b>Abstract .....</b>	<b>5</b>
<b>Acknowledgements .....</b>	<b>7</b>
<b>Table of Contents .....</b>	<b>8</b>
<b>List of Tables .....</b>	<b>14</b>
<b>List of Figures .....</b>	<b>20</b>
<b>List of Abbreviations.....</b>	<b>23</b>
<b>Chapter 1 Background and the hypotheses for research .....</b>	<b>26</b>
Abstract .....	26
1.1 Assessing professionalism and competence in doctors: an overview .....	27
1.1.1 The concept of professionalism .....	27
1.1.2 Professionalism in medicine.....	28
1.1.3 Competence <i>versus</i> expertise .....	30
1.1.4 The concept of certification .....	31
1.2 Ensuring the quality of assessments in medicine: rationale for the work .....	33
1.3 Establishing the validity of an assessment .....	34
1.3.1 Sources of evidence in a validity study .....	35
1.3.2 The predictive validity of medical assessments .....	36
1.4 Scope of the work.....	42
1.5 Aim and hypotheses .....	43
Summary .....	44
<b>Chapter 2 Specificity of the UK training system and description of the data sources .....</b>	<b>45</b>
Abstract .....	45
2.1 Medical Career Options: an overview of the medical education system and physician's training .....	46
2.2 Measures of Performance in the UK Setting.....	49
2.2.1 Assessing knowledge.....	50
2.2.2 Assessing clinical skills.....	52



2.2.3	Assessing professional attitudes .....	57
2.3	Summary of the Sources of Criterion measures .....	60
2.4	Description of the Sources of data.....	62
2.4.1	MRCP(UK) .....	62
2.4.2	MRCGP .....	75
2.4.3	Clinical Oncology specialist exam ('FRCR') .....	78
2.4.4	Specialty Certificate Exams ('SCEs') .....	80
2.4.5	Cardiology specialist exam: Knowledge Based Assessment ('CKBA') .....	81
2.4.6	Annual Review of Competence Progression ('ARCP') .....	82
2.4.7	List of Registered Medical Practitioners ('LRMP') .....	83
2.4.8	Investigation by the GMC (Fitness to Practice review documentation) .....	85
	Summary .....	86
<b>Chapter 3</b>	<b>Methodology.....</b>	<b>88</b>
	Abstract.....	88
3.1	Sample and sampling technique .....	88
3.2	Materials .....	90
3.2.1	MRCP (UK) Dataset details .....	90
3.2.2	MRCGP Dataset details.....	90
3.2.3	FRCR Dataset details.....	91
3.2.4	SCEs Datasets details.....	92
3.2.5	Cardiology Dataset details.....	93
3.2.6	ARCP Dataset details .....	93
3.2.7	LRMP Dataset details.....	93
3.2.8	Fitness to Practice Dataset obtained from the GMC.....	94
3.3	Criterion measures and Variables.....	94
3.3.1	Measures .....	95
3.3.2	Variables.....	98
3.4	Design.....	101
3.5	Missing Data.....	101

3.5.1	Systematic Missing Data Issues .....	101
3.5.2	Random missing data .....	102
3.5.3	Handling missing data .....	103
3.6	Procedure .....	103
3.7	Statistical Treatment .....	104
3.7.1	Correlation coefficients .....	104
3.7.2	Correction for attenuation (disattenuation) .....	104
3.7.3	Correction for range restriction .....	105
3.7.4	Comparison of means and analysis of variance .....	106
3.7.5	Linear regression .....	106
3.7.6	Chow test .....	107
3.7.7	Linear multi-level models .....	107
3.7.8	Logistic regression .....	108
3.7.9	Meta-analysis .....	109
3.7.10	Bootstrapping .....	110
3.7.11	Structural Equation Models .....	111
3.8	Statistical software .....	111
3.9	Limitations .....	112
3.10	Ethical Approval.....	113
	Summary .....	114
<b>Chapter 4</b>	<b>The MRCP(UK) internal prediction results.....</b>	<b>115</b>
	Abstract .....	115
4.1	Descriptive Statistics.....	116
4.2	MRCP(UK) as a Process.....	119
4.3	Learning Curves and the Meaning of the First Attempt Scores .....	120
4.4	Correlations Between first attempt scores at MRCP(UK) Part I, Part II and PACES ...	126
4.4.1	Between First Attempt Scores at MRCP(UK) Part I, Part II and PACES .....	126
4.4.2	Between First Attempt Results and Overall Number of Attempts .....	129
4.5	Contrasting Groups.....	129

4.5.1	MRCP(UK) Highfliers .....	129
4.5.2	Differences based on demographic characteristics .....	132
4.6	Regression Models .....	147
4.7	Structural Equation Model.....	150
	Summary and Discussion .....	152
<b>Chapter 5 Relationship between MRCP(UK) and Measures of Knowledge .....</b>		<b>156</b>
	Abstract.....	156
5.1	Specialty Certificate Exams .....	157
5.1.1	Descriptive statistics .....	157
5.1.2	Inferential statistics .....	163
5.2	Cardiology CKBA.....	174
5.2.1	Descriptive statistics .....	174
5.2.2	Inferential Statistics.....	175
5.3	FRCR – Clinical oncology .....	177
5.3.1	Descriptive statistics .....	178
5.3.2	Inferential statistics .....	185
5.4	MRCGP Applied Knowledge Test .....	193
5.4.1	Descriptive statistics .....	194
5.4.2	Inferential statistics .....	195
	Summary and Discussion .....	199
<b>Chapter 6 Assessment of Clinical Skills and Attitudes .....</b>		<b>203</b>
	Abstract.....	203
6.1	FRCR2 clinical examination .....	203
6.2	MRCGP Clinical Skills Assessment.....	205
6.2.1	Descriptive statistics .....	205
6.2.2	Inferential statistics .....	206
6.3	The Annual Review of Competence Progression ('ARCP') .....	210
6.3.1	Descriptive statistics .....	211
6.3.2	Inferential statistics .....	214

6.4 Registration status based on the List of Registered Medical Practitioners.....	217
6.4.1 Descriptive statistics.....	218
6.4.2 Inferential statistics.....	219
6.5 Being Subject to Investigation by the GMC Fitness to Practice Procedures .....	225
6.5.1 Descriptive statistics.....	225
6.5.2 Contrasting Groups .....	227
Summary and Discussion.....	227
<b>Chapter 7 Meta-analyses .....</b>	<b>232</b>
Abstract .....	232
7.1 Meta-analyses of coefficients associated with examinations.....	232
7.2 Meta-analyses of general underperformance effect sizes.....	237
Summary and Discussion.....	241
<b>Chapter 8 Discussion of the results and summary .....</b>	<b>244</b>
Abstract .....	244
8.1 Summary of the Results .....	245
8.1.1 Relationship between MRCP(UK) parts.....	245
8.1.2 Relationship between MRCP(UK) and knowledge exams .....	245
8.1.3 Relationship between MRCP(UK) and clinical skills and performance measures .....	247
8.1.4 Results of the meta-analyses .....	250
8.2 Limitations and Directions for Future Research.....	251
8.3 Meaning of the Findings and General Implications.....	254
Summary .....	257
<b>Bibliography.....</b>	<b>259</b>
<b>Appendix A Ethical Approval.....</b>	<b>287</b>
<b>Appendix B Systematic Review of the Literature .....</b>	<b>289</b>
<b>Appendix C Example PACES scenario.....</b>	<b>293</b>
<b>Appendix D Additional Tables and Graphs to Chapter 4 .....</b>	<b>295</b>
SECTION 4.1.....	295

SECTION 4.3 .....	296
<b>Appendix E Meta-analyses: Funnel Plots .....</b>	<b>302</b>

## List of Tables

Table 1. Sources of evidence in a validation process. ....	36
Table 2. Composition of Part I papers with respect to the fields of medicine. ....	67
Table 3. Composition of Part II paper with respect to the fields of medicine .....	70
Table 4. List of clinical skills assessed during PACES across stations with a minimum of points in each station required to pass. ....	74
Table 5. List of registration statuses on the LRMP with explanations. ....	84
Table 7. The overall number of records from the SCE files with the list of years of examinations taken into account. ....	92
Table 8. Frequency of LRMP registration status occurrences over the 4-year period of time for which the data was collected. ....	94
Table 9. Key variables used in the analyses in this research with the represented construct, criterion source, and level of measurement. ....	99
Table 10. Distribution parameters for the number of attempts in MRCP(UK) parts variables. ....	118
Table 11. Mean score at final attempt in Part I for groups distinguished based on the total number of attempts in Part I with <i>SE</i> , 95% confidence intervals, and numbers of valid cases. ....	124
Table 12. Distribution parameters for the first attempt results in the parts of MRCP(UK). ....	125
Table 13. Correlation coefficients (Spearman's $\rho$ , Pearson's $r$ and bootstrapped Pearson's $r$ ) for first attempt scores at MRCP(UK) parts. ....	126
Table 14. Correlation coefficients between MRCP(UK) parts with parameters required in the process of range derestriction and disattenuation .....	128
Table 15. Correlation coefficients (Spearman's $\rho$ , Pearson's $r$ and bootstrapped Pearson's $r$ ) between first attempt scores and total number of attempts in MRCP(UK) parts – a comparison. ....	129
Table 16. Comparison of mean scores between the Highfliers and Typical Candidates groups in MRCP(UK) parts (independent samples t-test results with bootstrapping and Mann-Whitney Z results) with effect sizes. ....	131

Table 17. Comparison of mean scores at MRCP(UK) parts between groups based on sex (independent samples t-test results, also bootstrapped, and Mann-Whitney Z results) with effect sizes.....	133
Table 18. Comparison of mean scores in MRCP(UK) parts between groups based on declared ethnicity (independent samples t-tests, also bootstrapped, and Mann-Whitney Z results) with effect sizes. ....	135
Table 19. Comparison of mean scores at MRCP(UK) parts between groups based on PMQ (independent samples t-tests, also bootstrapped, and Mann-Whitney Z results) with effect sizes.....	137
Table 20. Comparison of mean scores at MRCP(UK) parts between groups based on being a probable UK trainee (independent samples t-tests, also bootstrapped, and Mann-Whitney Z results) with effect sizes. ....	139
Table 21. Significant effects in a 2x2x2x2 factorial bootstrapped ANOVA on MRCP(UK) Part I scores with ethnicity, sex, PMQ and being a probable UK trainee as factors. ....	140
Table 22. Significant effect in a 2x2x2x2 factorial bootstrapped ANOVA on MRCP(UK) Part II scores with ethnicity, sex, PMQ and being a probable UK trainee as factors. ....	141
Table 23. Significant effects in a 2x2x2x2 factorial bootstrapped ANOVA on MRCP(UK) PACES scores with ethnicity, sex, PMQ and being a probable UK trainee as factors. ....	141
Table 24. Summary of the regression models fitted for Part I, Part II and PACES scores as dependent variables and demographic factors as predictors. ....	149
Table 25. Standardised and non-standardised coefficients with significance levels for the structural equation model fitted for MRCP(UK) scores and demographic factors.....	152
Table 26. Records of attempts and overall pass-rates by specialty in Specialty Certificate Exams. ....	158
Table 27. Demographic characteristics of the SCE candidates within each specialty (as per cent of total number of cases).....	159
Table 28. Distribution parameters for each SCE examination scores with one- sample K-S test results. ....	160
Table 29. Correlation coefficients (Pearson $r$ ) between MRCP(UK) parts scores and SCE scores. ....	164

Table 30. Correlation coefficients between MRCP(UK) results and SCE results, by SCE, for each part of MRCP(UK).....	166
Table 31. Comparison of mean scores between MRCP(UK) Highfliers and Typical Candidates for twelve specialties (independent samples t-test results) with effect sizes ( <i>r</i> ). ....	168
Table 32. Summary of the regression models fitted for the SCE scores with MRCP(UK) parts as predictors. ....	169
Table 33. Summary of the regression models fitted for the SCE scores with missing MRCP(UK) parts scores estimated with the EM algorithm. ....	171
Table 34. Summary of the general SCEs linear model with and without EM algorithm substituted missing values. ....	174
Table 35. Parameters for derestriction of range and disattenuation of coefficients between CKBA and MRCP(UK). ....	175
Table 36. Summary of the general SCEs & Cardiology linear model with missing values estimated using EM algorithm. ....	177
Table 37. Number of candidates who passed or failed FRCR1 modules with the number of attempts and pass-rates.....	179
Table 38. Counts of the Total Number of Attempts in FRCR1, overall and in division by the FRCR1 Rank groups.....	180
Table 39. Mean and median scores for Total Number of Attempts in FRCR1 by FRCR1 Rank. ....	180
Table 40. Comparison of the number of modules attempted and passed for groups of Dropouts and Failed. ....	181
Table 41. Performance in FRCR2: mean scores (with SD) and mean number of attempts in division by FRCR1 Rank groups.....	184
Table 42. Correlation coefficients (Pearson's <i>r</i> ) between MRCP(UK) and performance measures in FRCR1 and FRCR2. ....	186
Table 43. Correlation coefficients (Pearson's <i>r</i> ) between MRCP(UK) scores and scores in FRCR1 modules.....	187
Table 44. Derestriction of range and disattenuation of correlation coefficients between MRCP(UK) and selected FRCR performance measures with parameters required for corrections.....	188



Table 45. Mean scores (with SDs) in the MRCP(UK) parts in division by three FRCR1 Rank groups .....	189
Table 46. Summary of the regression models for selected FRCR performance measures as dependent variables . ....	191
Table 47. Summary of the regression models for selected FRCR performance measures with the EM algorithm imputed values. ....	192
Table 48. Summary of the logistic regression models summary for FRCR2 Pass/Fail score before and after imputing missing values using EM algorithm. ....	192
Table 49. Comparison between FRCR models and the SCEs & Cardiology joint model using Chow test. ....	193
Table 50. Cross-table for MRCP(UK) and MRCGP Pass/Fail outcomes. ....	194
Table 51. Pearson's product moment correlation coefficients between MRCP(UK) parts scores and AKT parts scores.....	195
Table 52. Range derestriction and disattenuation of the correlation coefficients between MRCP(UK) and AKT, with parameters required for both corrections.....	196
Table 53. Comparison of mean AKT scores between the MRCP(UK) Highfliers and Typical Candidates. ....	197
Table 54. Summary of the linear regression models for MRCGP AKT scores. ....	198
Table 55. Comparison of AKT, SCEs and FRCR1 models using Chow test. ....	199
Table 56. Summary of the FRCR2 clinical examination linear regression models before and after EM algorithm.....	205
Table 57. Correlation coefficients (Pearson's $r$ ) between MRCGP CSA and MRCP(UK) parts. ....	206
Table 58. Derestriction of range and disattenuation of the correlation coefficients between MRCP(UK) and CSA and the parameters required for both corrections. ....	207
Table 59. Comparison of mean scores in CSA between MRCP(UK) Highfliers and Typical candidates.....	208
Table 60. Linear regression models for CSA scores with MRCP(UK) parts as predictors.....	208
Table 61. Results of the Chow test comparisons between models for the three CSA scoring schemes. ....	209

Table 62. Comparisons between CSA Equated regression model and AKT, SCEs with Cardiology and FRCR1 models using the Chow test.....	210
Table 63. Comparisons of mean scores in MRCP(UK) parts (independent samples t-tests) with Overall Progress Assessments as the factor.....	214
Table 64. Comparisons of mean score in MRCP(UK) with Overall Progress Assessments as the factor for groups with different number of assessments.....	216
Table 65. Summary of the logistic regression model for ARCP Unsatisfactory Progress as dependent variable with MRCP(UK) parts and demographic characteristics as predictors.	217
Table 66. Comparison of mean scores in MRCP(UK) parts (independent samples t-tests) for Licence Issues as the factor. ....	220
Table 67. Comparison of mean scores in MRCP(UK) (independent samples t-tests) between doctors with relinquished licences and the rest of the sample. ....	221
Table 68. Comparison of mean scores in MRCP(UK) parts (independent samples t-tests) results between doctors erased for administrative reasons and not erased. ....	222
Table 69. Summary of the logistic regression model for occurrences of Licence Issues as dependent binary variable with MRCP(UK) parts and demographic characteristics as predictors. ....	223
Table 70. Summary of the logistic regression model for Voluntary Erasure as binary dependent variable with MRCP(UK) parts and demographic characteristics as predictors.	224
Table 71. Summary of the logistic regression model for Administrative Erasures as binary dependent variable with MRCP(UK) parts and demographic characteristics as predictors.	225
Table 72. Overlap between GMC FtP list of investigated cases and groups of doctors with selected LRMP registration statuses. ....	226
Table 74. Summary of three meta-analytical models for coefficients between Part I, Part II and PACES and the criteria associated with examinations. ....	234
Table 73. Raw and corrected Pearson's r coefficients (with SEs) between MRCP(UK) parts and criterion measures used in meta-analyses (with sample sizes). ....	235
Table 75. Effect sizes for underperformance related criteria for all MRCP(UK) parts (with SE and degrees of freedom).....	238
Table 76. Summary of the meta-analytical models for measures of underperformance related to the three MRCP(UK) parts. ....	239

Table B1. List of publications from a systematic review on psychometric properties of medical education related tests. ....	289
Table D1. Frequencies of passed and failed MRCP(UK) Part I candidates based on the number of attempts.....	295
Table D2. Frequencies of passed and failed MRCP(UK) Part II candidates based on the number of attempts.....	295
Table D3. Frequencies of passed and failed MRCP(UK) PACES candidates based on the number of attempts.....	296
Table D4. Mean scores for the twelve groups based on the Total Number of Attempts in Part I (showing the learning process). ....	298
Table D5. Mean scores for the twelve groups based on the Total Number of Attempts in Part II (showing the learning process). ....	299
Table D6. Mean scores for the twelve groups based on the Total Number of Attempts in PACES (showing the learning process).....	300
Table D7. Results of the post-hoc REGW Q test for one-way ANOVA on first attempt Part II scores with the twelve groups based on the Total Number of Attempts being the factor. ....	301
Table D8. Results of the post-hoc REGW Q test for one-way ANOVA on first attempt Part II scores with the twelve groups based on the Total Number of Attempts being the factor. ....	301

## List of Figures

Figure 1. A simplified overview of typical postgraduate medical training paths in the UK. ..	46
Figure 2. Division of responsibilities between major institutions influencing the training of a doctor in the United Kingdom.....	48
Figure 3. Description of potential postgraduate examinations that can be taken by doctors in the United Kingdom. ....	51
Figure 4. Possible and secured sources of criteria for the predictive validity study of MRCP(UK). ....	61
Figure 5. Explanation of the order in which MRCP(UK) needed to be attempted.....	66
Figure 6. Diagram of current PACES station descriptions and rotation scheme.....	73
Figure 7. Possible outcomes and strategies in attempting FRCR1. ....	79
Figure 8. Timeline of key stages (mean number of weeks) of the MRCP(UK) process, from PMQ to passing MRCP(UK). Number of valid cases depended on the stage (due to censoring) from n=11,990 to n=35,950 candidates . ....	116
Figure 9. MRCP(UK) as a selective process: numbers of candidates after each stage of attempting the exam (based on n=39,335).....	119
Figure 10. Mean scores in Part I per attempt for groups based on total number of attempts – approximation of the learning curve.....	121
Figure 11. Visualised mean bootstrapped results (with standard error) in MRCP(UK) parts for the Typical Candidates and Highfliers groups (number of valid cases provided in Table 16).....	130
Figure 12. Mean scores (with SE) in Part I and Part II for groups based on sex and ethnicity (interaction effect). ....	142
Figure 13. Mean scores (with SE) in Part I for groups based on PMQ and ethnicity (interaction effect). ....	143
Figure 14. Mean scores (with SE) in all MRCP(UK) parts for groups based on PMQ and probable training (interaction effect). ....	144
Figure 15. Mean scores (with SE) in PACES for groups based on sex and probable UK training (interaction effect). ....	145
Figure 16. Mean scores (with SE) in Part II for groups based on ethnicity, PMQ and probable UK training (interaction effect). ....	146

Figure 17. Structural Equation Model describing relationships between MRCP(UK) first attempt scores and demographic factors (sex, ethnicity, PMQ, and being a probable UK trainee); n= 50,311.....	151
Figure 18. Comparison of mean Part II Z-scores (with SE) between specialties.....	162
Figure 19. Comparison of mean PACES Z-scores (with SE) between specialties. ....	162
Figure 20. Fitted separate regression lines for SCEs with aggregated MRCP(UK) Z-scores. ....	172
Figure 21. Summary of the multi-level model for the SCE scores with MRCP(UK) parts as predictors (MLwiN output). ....	173
Figure 22. Summary of the multi-level model for the SCEs with Cardiology with MRCP(UK) parts as predictors (MLwiN output).....	177
Figure 23. Comparison of first attempt Mean Module Marks (with 95% CI) between FRCR1 Rank groups. ....	182
Figure 24. Distribution of the Total Number of Attempts in FRCR2 by candidates who passed and failed FRCR2. ....	183
Figure 25. Distribution of FRCR2 written component standardised scores. ....	185
Figure 26. Frequency of the number of assessments in the analysed ARCP sample (n=2,979). ....	212
Figure 27. Frequencies of the number of assessments in groups based on the Overall Progress Assessment (All satisfactory versus At least one unsatisfactory).....	213
Figure 28. Standard forest plots for effects of MRCP(UK) parts on examinations criteria..	236
Figure 29. Funnel plot for meta-analytical model on coefficients associated with Part I and examinations.....	237
Figure 30. Standard forest plots for effects of MRCP(UK) parts on underperformance criteria.....	240
Figure D1. Mean scores in Part II per attempt for groups based on total number of attempts – approximation of the learning curve. ....	296
Figure D2. Mean scores in PACES per attempt for groups based on total number of attempts – approximation of the learning curve .....	297
Figure E1. Funnel plot for meta-analytical model on coefficients associated with Part II and examinations.....	302

Figure E2. Funnel plot for meta-analytical model on coefficients associated with PACES and examinations. ....	302
Figure E3. Funnel plot for meta-analytical model on coefficients associated with Part I and underperformance criteria.....	303
Figure E4. Funnel plot for meta-analytical model on coefficients associated with Part II and underperformance criteria.....	303
Figure E5. Funnel plot for meta-analytical model on coefficients associated with PACES and underperformance criteria.....	304

## List of Abbreviations

(in an alphabetical order)

<b>AERA</b>	American Educational Research Association
<b>AKT</b>	Applied Knowledge Test (written part of MRCGP)
<b>AKT CM</b>	AKT Clinical Medicine (partial score in AKT)
<b>AKT EI</b>	AKT Evidence Investigation (partial score in AKT)
<b>AKT QO</b>	AKT Organisational Questions (partial score in AKT)
<b>APA</b>	American Psychological Association
<b>ARCP</b>	Annual Review of Competence Progression
<b>BAPIO</b>	British Association of Physicians of Indian Origin
<b>BCS</b>	British Cardiovascular Society
<b>BMAT</b>	BioMedical Admissions Test
<b>CI</b>	Confidence Interval
<b>CCT</b>	Certificate of Completion of Training
<b>CHRE</b>	Council of Healthcare Regulatory Excellence
<b>CKBA</b>	British Cardiovascular Society Knowledge Based Assessment
<b>CMT</b>	Core Medical Training
<b>CPD</b>	Continuing Professional Development
<b>CSA</b>	Clinical Skills Assessment (clinical part of MRCGP)
<b>EM</b>	Estimation-Maximization algorithm
<b>FRCR</b>	Fellowship of the Royal College of Radiologists' in Clinical Oncology
<b>FRCR1</b>	First FRCR exam
<b>FRCR2</b>	Final FRCR exam
<b>FTSTA</b>	Fixed Term Specialty Training Appointments
<b>GMC</b>	The General Medical Council
<b>GMC FtP</b>	GMC Fitness to Practise (panel or procedures)
<b>GMC Number</b>	Individual number assigned to each registered doctor by the GMC
<b>GCSEs</b>	General Certificates of Secondary Education
<b>HSCIC</b>	Health & Social Care Information Centre
<b>IELTS</b>	International English Language Testing System
<b>IMGs</b>	International Medical Graduates
<b>IRT</b>	Item Response Theory
<b>JRCPTB</b>	Joint Royal Colleges of Physicians Training Board

<b>K-S test</b>	One-sample Kolmogorov-Smirnov test
<b>LAT</b>	Locum Appointment Training
<b>LETBs</b>	Local Educational and Training Boards
<b>LRMP</b>	List of Registered Medical Practitioners
<b>LSA</b>	Trust Doctor or Locum for Service doctor
<b>MAR</b>	Missing At Random
<b>MCAR</b>	Missing Completely at Random
<b>MCAT</b>	Medical College Admissions Test
<b>MCCQE</b>	Medical Council of Canada Qualifying Examination
<b>MCQs</b>	Multiple-Choice Questions
<b>MPTS</b>	Medical Practitioners Tribunal Service
<b>MRCGP</b>	Examination for the Membership of the Royal College of General Practitioners
<b>MRCP(UK)</b>	Examination for the Membership of The Royal Colleges of Physicians of the United Kingdom
<b>NCLEX</b>	National Council Licensure Examination
<b>NCME</b>	National Council on Measurement in Education
<b>NPSA</b>	National Patient Safety Agency.
<b>OSCE</b>	Objective structured clinical examination
<b>PACES</b>	Standardised clinical assessment within MRCP(UK)
<b>Part I</b>	Part I of the MRCP(UK) examination (written)
<b>Part II</b>	Part II of the MRCP(UK) examination (written)
<b>PLAB</b>	Professional and Linguistic Assessments Board examination
<b>PMQ</b>	Primary Medical Qualification
<b>QLEX</b>	Quebec Licensing Examination
<b>RCGP</b>	Royal College of General Practitioners
<b>REML</b>	Restricted Maximum Likelihood estimation method
<b>RCP</b>	Royal Colleges of Physicians of the United Kingdom
<b>RCP ID</b>	Individual number assigned to each candidate by the RCP
<b>RCR</b>	Royal College of Radiologists
<b>REGW</b>	Ryan-Einot-Gabriel-Welsch (R-E-G-W) Multiple Stepdown Procedure (post-hoc test)
<b>RMSEA</b>	A root mean square of approximation statistic (used in SEMo)
<b>SATs</b>	United States Scholastic Aptitude Tests



<b>SCE or SCEs</b>	Specialty Certificate Exam(s)
<b>SD</b>	Standard Deviation
<b>SE</b>	Standard Error (of Means, of Estimate, as appropriate)
<b>SEM</b>	Standard Error of Measurement
<b>SEMo</b>	Structural Equation Modelling
<b>SHAs</b>	Strategic Health Authorities
<b>SD</b>	Standard deviation
<b>ST</b>	Higher Specialty Training
<b>Standards</b>	The Standards of Psychological and Educational Testing, published by APA, AERA and NCME
<b>TAB</b>	Team Assessment of Behaviour, part of WBA
<b>TLI</b>	Tucker-Lewis Index
<b>UK</b>	The United Kingdom
<b>US</b>	The United States of America
<b>USMLE</b>	United States Medical Licensing Examination
<b>VIF</b>	Variance Inflation Factor
<b>WBAs</b>	Workplace Based Assessments

## Chapter 1

### Background and the hypotheses for research

#### ABSTRACT

*Reports of a decrease in public trust towards the healthcare system and medical profession have been present in the literature for several decades. As a means of addressing public concern, a New Professionalism framework was proposed. This affected all activities of the medical profession, including medical education and examinations. The three key pillars of professionalism in medicine were identified as knowledge, skills, and professional attitudes, and have subsequently been embedded in the medical curriculum. The same three concepts were incorporated into the design of the MRCP(UK) examination. The general purpose of this research was to address the question of whether MRCP(UK) - as one of the key medical exams in the career of a UK physician - truly assesses those three components. One way of approaching this subject was through a predictive validity study, which was the key focus of this research. A literature review provided examples of predictive validity studies of other high-stakes medical exams with a thorough description of the employed methods. Based on the literature review it was hypothesised that MRCP(UK) would predict subsequent knowledge exams, clinical skills assessments, and on the job performance.*

In September 2014, approximately<sup>1</sup> 267,500 doctors were registered with a licence to practice in the United Kingdom (GMC, 2014a). Of those, almost 82,000 were specialist doctors, amongst which nearly 14,000 were physicians working as consultants or specialty registrars in hospitals (The Federation of the Royal Colleges of Physicians of the United Kingdom ('RCP'), 2011). The official Hospital Episodes Statistics for the years 2012-13 recorded approximately 4.4 million episodes requiring consultation by a physician in England only (Health & Social Care Information Centre 'HSCIC', 2013a). Therefore, the demand for medical services was and is immense, and in order to provide effective care to the patients requiring it, doctors must be professional and competent. Crucially, they need to be able to demonstrate professionalism and competence to their patients, their colleagues, their employers, and the medical regulator. One important way in which

---

<sup>1</sup> The actual number of doctors registered with the GMC changes daily due to multiple factors. The numbers provided above are to demonstrate the potential scale of the issue of providing high quality medical services to the society.

doctors can demonstrate this is by passing high-stakes assessments. This PhD set out to test the quality of one such assessment, the examination for the Membership of the Royal Colleges of Physicians of the United Kingdom ('MRCP(UK)'), which is taken by approximately 6,000 doctors around the world every year.

## **1.1 ASSESSING PROFESSIONALISM AND COMPETENCE IN DOCTORS: AN OVERVIEW**

Professionalism and competence are qualities that have been widely discussed in the literature; however, they are still considered hard to define. One possible reason for this might be that they are two intertwined concepts that simultaneously encompass certain behaviours and attitudes that encourage public trust (Evetts, 2006; Hodges *et al.*, 2011; Pellegrino, 2002; Svensson, 2006). Another possible reason is that competence, a concept considered hard to grasp (Shanteau, 1992; Stoof, Martens, van Merrienboer, & Bastiaens, 2002), is often regarded as the cornerstone of professionalism (Arnold & Stern, 2006; Pellegrino, 2002). Svensson (2006) demonstrated that knowledge, competence, and skills are the most commonly used synonyms for professionalism, even though they are different concepts, and according to Eraut the prevailing belief is that "the professionals [...] know what competence is and do not need to spell it out" (Eraut, 2003, p. 116).

### **1.1.1 The concept of professionalism**

'Professionalism' has been discussed from sociological, operational, interpersonal, and personal approaches (Hodges *et al.*, 2011; Martimianakis, Maniate, & Hodges, 2009) on several levels: from being a role or social construction, to being a means of social control. A 'profession' is commonly defined in the literature as a form of community based on occupation that shares a particular ethical code (Durkheim, 1992 in Evetts, 2003); hence, professions control expertise. Experts are required so that the members of the society can be protected against incompetence and exploitation in areas where they do not possess enough knowledge to evaluate the quality of service themselves (Eraut, 2003, p. 1). Eraut (2003) further suggests that professionalism is an ideology in which knowledge has a supreme role. Professions are, therefore, entities formed by a group of experts in a field, who ensure the control over the quality of services provided and impose a code of conduct. In return for high-quality services, a profession – in particular the medical profession – receives trust from society (Chamberlain, 2010; Cohen, 2006; Irvine, 1999; Landon, Normand, Blumenthal, & Daley, 2003; RCP, 2005; Stevens, 2001). This trust manifests as a lack of interference in profession related issues, such as setting standards of performance, choosing methods for training, and admitting new members to the profession (Cohen,

2006; Mechanic, 1996). These privileges and their high social status give the professions social influence and power and grant them a monopoly over their own knowledge base.

Professions decide which knowledge is transferred to the new members. They influence higher education curricula (Eraut, 2003), shape post-graduate educational programmes, and are responsible for the continuing professional development of their members. However, the knowledge that is passed between the members of a community comprises more than purely factual knowledge; it also encompasses tacit knowledge or a 'hidden curriculum', regarded as exposure to the specific environment that reinforces attitudes and behaviours deemed desirable by the profession (Hafferty & Franks, 1994; Martimianakis *et al.*, 2009). Some researchers argue that tacit knowledge is key for developing professional behaviour and shaping core values (Cottingham *et al.*, 2008; Suchman *et al.*, 2004); however, more recently efforts have been made to teach professionalism and thus influence the hidden curriculum (Goldstein *et al.*, 2006).

The power of the professions and their exclusivity made them vulnerable to accusations of being self-serving rather than altruistic, and of using their privileged position in society to control the service market (Siegrist, 1994). With an increasing access to media and information, cases of malpractice or fraudulent behaviours have larger impacts on perception of the professions, which makes them only as credible as their weakest members (Eraut, 2003, p. 117) and further undermines their position. The increase in public awareness and concern over professional competence and professional monopolies encourages governmental interventions and supervision. Criticisms of the professions are not without merit, as professional foundations lie in being service-centred rather than client-centred (Eraut, 2003, p. 4), while the latter is a standard for the majority of market economy service providers.

### **1.1.2 Professionalism in medicine**

In the field of medicine, however, a shift in focus on client service has also been observed (Gill & Griffin, 2010), which was reflected in the changing language of the *Good Medical Practice* guidelines published by the General Medical Council ('GMC'). This corresponds to an increase in patient awareness indicated by a raising number of complaints in recent years (Archer, Regan de Bere, Bryce, & Nunn, 2014; NPSA, 2011), increased criticism of the medical profession (Allsop, 2006), and a decline in public trust (Jacobs, 2005; Mechanic & Schlesinger, 1996; Mechanic, 1996; Pfadenhauer, 2006; Schlesinger, 2002; Stevens, 2001). The perceived erosion of public trust led to the Medical Practitioners Tribunal Service

(‘MPTS’) being established in 2012. Operationally separate from the GMC, but accountable to Parliament and to the GMC Council, it makes decisions about doctors’ fitness to practise that are separate from the GMC’s investigations. GMC Chief Executive Niall Dickson called the launch of the MPTS “the biggest change to doctors’ fitness to practise hearings for more than 150 years”, explaining how it aims to “strengthen professional and public confidence that our hearings are impartial, fair and transparent – the fact that the service is led by a judicial figure who has a direct line to Parliament should provide that assurance” (MPTS, 2012).

The issue of eroding trust is country-specific and hence dependent on the particulars of the healthcare system. In the United States, one study that looked into the decline in trust was conducted by Blendon and colleagues (Blendon, Benson, & Hero, 2014). They compared US patients’ perspectives with the views of patients’ from other industrialised countries that participated in the International Social Survey Programme (years 2011-2013). Although in this survey the United States was indeed ranked low, which would be consistent with the reports of American authors such as Cohen (2006), the UK was ranked fourth from the top, with 76% participants agreeing that ‘all things considered doctors can be trusted’, and was ranked seventh in terms of overall satisfaction with the last treatment received. The 2014 annual Ipsos Mori Trust Poll reported that doctors in the UK are (still) the most trusted profession compared to teachers, scientists, judges, priests, police among others (Ipsos Mori, 2015), and 90% of interviewees agreed that doctors tell the truth. Calnan and Sanford looked into the discrepancy between the perceived trustworthiness of doctors and evidence of a decline in public satisfaction with the UK healthcare system during the past 20 years (Calnan & Sanford, 2004). Their original research confirmed the Ipsos Mori poll results; trust in medical professionals was high, but trust in the health service managers responsible for organisation and finance was low. Therefore, it seems that the decline in trust may not be a straightforward issue, and may be system- rather than profession-related.

The functioning of the healthcare system and public expectations continue to pose challenges to the medical profession. Irvine (2001) and others (Norcini & Talati, 2009; Shaw, Cassel, Black, & Levinson, 2009) argue that the extended regulatory framework from the government and the NHS was introduced due to a combination of high-profile cases of medical malpractice and perceived lack of accountability and maintained self-focus of the medical profession. It resulted in the adoption of the concept of the New Professionalism (Hargreaves, 1994) in medicine. The framework of New Professionalism was designed as a

set of qualities agreed between the public and the medical profession on what it means to be a good doctor. In practical terms, the adoption of the New Professionalism resulted in new standards of medical practice set out in *Good Medical Practice* (GMC, 1995) which replaced *The Blue Book* (GMC, 1963), and in implementation of those standards into the structure of the profession. At an operational level, the adoption of the new standards resulted in the Fitness to Practice procedures (1997) and revalidation (GMC, 2013c), but also in changes in the medical curricula (Irvine, 2001). At an individual level the New Professionalism has been identified as an increased dedication to quality and professional standards (Jones & Green, 2006).

Although the conditions in which professions are set and operate are knowledge-, culture-, and time-specific (Hodges *et al.*, 2011), several researchers ventured to investigate the generic characteristics of a professional (Hickson & Thomas, 1969; Millerson, 1964). According to these studies, such features included, among other things, knowledge, altruistic service, ethics, skilfulness, loyalty, being impartial, etc. By modern definitions, professionalism in medicine refers to ethical aspects or personal qualities of a doctor (Hilton & Slotnick, 2005; Hodges *et al.*, 2011; Swick, 2000), such as humanism, accountability, altruism, reflectiveness, striving for excellence (Arnold & Stern, 2006), or conscientiousness (Finn, Sawdon, Clipsham, & McLachlan, 2009). In terms of medical philosophy professionalism is a virtue-based concept, and it has been equated with thorough understanding of principles of ethics, such as benevolence, fidelity to trust, truthfulness, intellectual honesty, compassion, and courage (Pellegrino, 2002). These aspects of morality reflect on all actions, duties, and events in a life of a doctor, and as such they create an entity called professionalism (Pellegrino, 2002). With reference to the above-mentioned virtues, professionalism is a form of social contract (Cruess, Cruess, & Johnston, 2000) that requires a commitment from a physician to possess the necessary knowledge and skills (a competence) to help a patient, and to use this competence in the best interest of that patient (Pellegrino, 2002).

### **1.1.3 Competence *versus* expertise**

Competence is a more narrow term than expertise, perceived as a combination of knowledge and skills (Fernandez *et al.*, 2012), or sometimes as a combination of skills, knowledge and attitudes (Stoof *et al.*, 2002) that sets a standard of performance for a particular profession (Maudsley & Strivens, 2000). However, this standard is in itself problematic, as the literature is somewhat ambiguous when competence and expertise are concerned (Herling, 2000). Based on the review by Fernandez and colleagues (Fernandez *et*

*al.*, 2012), who analysed published articles in the field of medical education, medical educators would often call for excellence or expertise when referring to the aspects of competence. This is shown in the definitions presented in the Fernandez *et al.* paper. Some of them refer to integrating complex data, acting under uncertainty, or managing ambiguous problems, which are the key aspects of being an expert. An example of a publication that explicitly refers to excellence is the *Tooke Report* (Tooke *et al.*, 2008), which underlines pursuit to excellence as a major principle for postgraduate training. Other examples come from Prof. Sir Kenneth Calman and his colleagues who have written that “Patients are entitled to expect safe and effective care and treatment by staff who are *expert* in what they do” (Calman, Temple, Naysmith, Cairncross, & Bennett, 1999, p.33), and from an ongoing debate over the standard demanded of medical trainees and educators at various levels of medical education (Holmboe, Sherbino, Long, Swing, & Frank, 2010; Lee *et al.*, 2008; Smith & Greaves, 2010). In light of the common confusion over the use of the two concepts, it has been proposed that competence should correspond to the minimal level of efficiency required for a successful performance in a task, while expertise should refer to the optimal or high level of such efficiency (Herling, 2000).

As medical standards aim towards excellence, there is an extensive amount of literature on expert performance of physicians and other medical professionals (Chi, 2006; Epstein, 2002; Feltovich, Prietula, & Ericsson, 2006; Landon *et al.*, 2003; Norman, Eva, Brooks, & Hamstra, 2006; Shanteau, 1992). For example, it has been established that experts differ from novices in the extent of knowledge possessed and its depth, but also in terms of organization of that knowledge, and their ability to integrate new information with previous knowledge (Chi, 2006). Some researchers, such as Dreyfus and Dreyfus (1986), name tacit knowledge and intuition as necessary components of expert performance. Although studies on expertise vary depending on their methodology and focus, they usually underline two factors: the level of aggregation of domain specific knowledge, and accumulation of skills, otherwise referred to as skilled behaviour (Eraut, 2003, p. 110)). This confirms the notion that those two components constitute the foundation of professional development, as discussed by Pellegrino (2002) and Arnold & Stern (2006).

#### **1.1.4 The concept of certification**

As professions have historically been responsible for training and education of their members, they have also developed licencing procedures (Eraut, 2003; Pfadenhauer, 2006), i.e. certification. Certification was designed to prove legitimacy and ability to perform certain duties in line with the code of conduct of the profession. This happens through

exams, which serve as means of assessing the competence of a professional. The forms of these exams vary across countries, but regardless of their local specificity, the general purpose of certification assessments is to guarantee that those who pass them have the qualities sufficient to perform safely as independent practitioners (Epstein, 2002; van der Vleuten, 2000; Wenghofer *et al.*, 2009). In the United Kingdom the key assessment for physicians is the examination for the Membership of the Royal Colleges of Physicians of the United Kingdom, or MRCP(UK), which is a qualification required to practice hospital medicine. The purpose of MRCP(UK) is to identify doctors who have “acquired necessary professional knowledge, skills, and attitudes, as defined in the published syllabus of the General Internal Medicine Curriculum, to enable them to benefit from a programme of higher specialist training with confidence and enthusiasm” (RCP, 2011). This statement clearly refers to professionalism, and defines not only the level of competence required from the MRCP(UK) candidates, but also frames the scope of qualities that the exam aims to assess, and therefore, outlines the psychometric constructs behind it.

Membership of the Royal Colleges of Physicians of the United Kingdom is a mandatory part of the educational curriculum for physicians and allows entry to higher specialist training. Therefore, it is an important part of the medical education process and a key step in a medical career in the UK. It is taken annually by approximately 6,000 doctors, comprising approximately 30% of all UK medical graduates, plus doctors trained outside the UK, all of whom either wish to practice in the UK or wish to improve their opportunities in their home country. MRCP(UK) is a three-stage exam administered by the Federation of the Royal Colleges of Physicians of London, Edinburgh, and Glasgow (jointly ‘RCPs’). It consists of two written exams (‘Part I’ and ‘Part II’) and one clinical exam (‘PACES’). A fuller description of this exam is provided in Chapter 2. Candidate fees for each part – set to cover the expenses of administering the exam – are substantial. MRCP(UK) is usually first attempted at the completion of Year 2 of the Foundation Programme. However, the formal requirement according to the RCPs regulations (RCP, 2011) is that MRCP(UK) may be taken at least 12 months from graduation, or after completion of the first year of the Foundation Programme. Completion of MRCP(UK) usually takes approximately 2 to 3 years; however, the RCP regulations provide a 7-year time window for passing all three parts. Otherwise the candidates need to repeat the process from the beginning. Notwithstanding the financial aspect of attempting the MRCP(UK), the exam certainly requires a high level of commitment from those taking it, which adds to its perception as a high-stakes exam.



## **1.2 ENSURING THE QUALITY OF ASSESSMENTS IN MEDICINE: RATIONALE FOR THE WORK**

The fact that the MRCP(UK) exam is high-stakes does not only stem from the personal or financial commitment of doctors who attempt it. The high-stakes nature of this exam extends from the concept of public trust in experts. As a major certification exam in the UK, the MRCP(UK) exam assesses competencies that translate into quality of medical services, and therefore, into public perception of physicians and collectively of the medical profession. For these reasons, evidence for the quality of the MRCP(UK) exam should be collected, challenged, and the outcomes should be reported and become public knowledge, which is what this PhD research to a certain extent set out to do.

In order to be considered high quality, any exam needs to meet certain psychometric evaluation criteria. Such criteria were agreed upon and published in the Standards of Educational and Psychological Testing (AERA, APA, & NCME, 2004) (the 'Standards'). The Standards require that a test is not only reliable, normalised, and standardised, but also valid. Validity supplies evidence of credibility (Norcini & Talati, 2009; Norcini *et al.*, 2011) of an examination, and provides significant support for the inferences made based on its outcomes. Establishing evidence that MRCP(UK) and other certification and qualification exams are of high quality also helps address recurring issues with the perceived accountability and level of self-regulation of the medical profession, both of which have come under scrutiny in recent years (Irvine, 1999; Pfadenhauer, 2006).

Apart from aiding the process of regaining public goodwill (Billington & Taylor, 2008; Dauphinee, 2005), the evidence of validity of medical exams also simply indicates that exams are good performance measures (Landon *et al.*, 2003). As such, evidence of validity provides justification for maintaining examinations (Cizek, 2012), and therefore, it may supply confirmation that the medical profession uses effective means to minimise medical errors (Epstein, 2002). Establishing evidence of appropriateness of selection procedures for the medical profession – such as MRCP(UK) – is necessary to demonstrate the competence of medical professionals. Validity is also essential to provide the means to address criticisms or any dispute over professional competence (Epstein, 2002; RCP, 2005). Therefore, in summary, confirmation of the validity of a medical exam or any other assessment (Van der Vleuten *et al.*, 2012) constitutes important evidence when the professionalism of medics is questioned.

### 1.3 ESTABLISHING THE VALIDITY OF AN ASSESSMENT

The importance of the concept of validity makes it a constantly developing term with a variety of definitions (Cizek, 2012; Kane, 2001; Lissitz & Samuelsen, 2007) and an evolving meaning. Initially, validity was regarded both as a feature of a test and a process that verified whether a test measures what it was intended to measure (Oerlik, 2002; Fredricksen, Mislevy, & Bejar, 1993; Gulliksen, 1950a; Kaplan & Saccuzzo, 1993). In a sense, validity was at the time perceived as a reverse process to that of designing a test. Over the course of validation procedures the results of a test would be set against the theory-induced expectations of what the results should have been, and the congruence between the two would imply validity. As there are potentially many methods of testing if such congruency exists, the indirect result of this approach towards establishing validity evidence was the sub-division of the term validity into several types. The names of these sub-types of validity specified the character of the method used, and therefore, the literature refers to e.g. construct validity, concurrent validity, contents validity, etc. (Anastasi & Urbina, 1997; Guion, 1980; Kane, 2001). This resulted in a multiplication of terminology and a structural expansion of validity, thus complicating the process of validation. It became a multi-stage and a multi-faceted process employing a variety of validity types, depending on the subjective perspective of a test designer.

In opposition to expanding the validity structure, and as a result of concerns about the opportunistic choice of validity evidence, a different approach to validity was proposed. Cronbach suggested that the rationale for the use of particular evidence type should be made based on a set of axioms and hypotheses resulting from theory (Kane, 2001) rather than based on convenience. This established the foundations for the concept of validity as a “unified and integrated evaluation of an interpretation” (Kane, 2001, p. 329). Validity started to refer to the degree to which the interpretations of test results could be advocated by theory, by associated evidence, and by original proposals of applications for that test. This approach is described in the Standards and currently supported by many prominent authors (AERA *et al.*, 2004; Anastasi & Urbina, 1997; Cronbach, 1990; Downing, 2003; Kane, 2001; Markus, 1998). Consequently, it has been proposed that it is not the test that should be validated, but a specific interpretation of its results. Hood (2009) argues that from a scientific realism point of view, both classic and modern approaches are logically equivalent; however, the Standards mostly avoid traditional nomenclature of validity types and instead propose to associate validity with the sources of its evidence.

### 1.3.1 Sources of evidence in a validity study

A typical validity study requires a pursuit of evidence that would support the interpretation of results of a particular test. Potential sources of evidence have been provided by the Standards, and are also summarized by Downing (2003). They are divided into five groups, as presented in Table 1.

The first group of evidence refers to the *contents* of a test, and includes for example, the representativeness of test items to the tested domain, the relationship between the domain and its contents, the quality of test questions, etc. The second group refers to *the response process*, such as familiarity of the format, quality control of scoring, or response key validation. The third group of evidence refers to the *internal structure* of a test. This class involves all item analyses irrespective of whether they are based on classic test theory or item response theory ('IRT'), and in particular, names reliability as one of the sources of validity evidence. This stems from the fact that reliability provides information on the accuracy of the results of a test, and without accuracy the results are not interpretable and, therefore, not valid. Hence, reliability constitutes the upper bound to validity, meaning that a test cannot be more valid than it is reliable. As posed by Downing, reliability is therefore "a necessary but not sufficient condition for validity" (Downing, 2004, p. 1007). This view was challenged by Moss (1994), but further counter argued in separate papers by Li (2003) and Mislevy (2004).

The fourth group of evidence relates to *the consequences of testing*, as proposed by Messick (1980), to remind all testers of ethical aspects of testing and justification for using tests. This group of evidence contains evaluation of potential consequences on the lives of tested subjects after e.g. adopting a particular method of setting a pass rate or pass-mark, or a particular test form. Although it has been included in the Standards, there is a certain level of controversy regarding the concept of consequences of testing. Cizek and colleagues (2010) noted that no consensus has been achieved with regards to what the role of this type of evidence in the validation process should be. The final, fifth, group of evidence comprises *relationships with other variables*, which inherently involves both convergent and discriminant correlations with external criteria, and as such it includes predictive validity coefficients.

**Table 1. Sources of evidence in a validation process.**

<i>Group 1 Contents of the test</i>	<i>Group 2 Response process</i>	<i>Group3 Internal structure</i>	<i>Group 4 Consequences of testing</i>	<i>Group 5 Relationship to other measures</i>
<ul style="list-style-type: none"> <li>• Test representativeness</li> <li>• Representativeness of items to the domain</li> <li>• Quality of test questions</li> <li>• Item writer qualifications</li> <li>• Other</li> </ul>	<ul style="list-style-type: none"> <li>• Response process</li> <li>• Familiarity of format</li> <li>• Quality of test marking devices</li> <li>• Key validation</li> <li>• Quality control of final scores</li> <li>• Pass/Fail decision</li> <li>• Reporting of the results to the candidates</li> <li>• Other</li> </ul>	<ul style="list-style-type: none"> <li>• Item difficulty and discrimination</li> <li>• Item Characteristic Curves</li> <li>• Inter-item and item-total correlations</li> <li>• Reliability</li> <li>• Standard errors of measurement</li> <li>• Other</li> </ul>	<ul style="list-style-type: none"> <li>• Evaluation of the consequences on lives of tested subjects</li> <li>• Pass/Fail rate establishing method</li> <li>• Decision on the test form</li> <li>• Other</li> </ul>	<ul style="list-style-type: none"> <li>• Correlations with other external variables both convergent and divergent</li> <li>• Test-criterion correlations</li> <li>• Generalisability of evidence</li> <li>• Other</li> </ul>

Source: Downing, 2003

### 1.3.2 The predictive validity of medical assessments

Predictive validity is based upon measuring the “effectiveness of a test in predicting an individual’s performance in specified activities” (Anastasi & Urbina, 1997, p. 119) over a specific period of time after the test. The time factor differentiates predictive validity from concurrent validity evidence, as the latter uses concurrent data. Although Anastasi and Urbina (1997) argue that there is a certain level of ambiguity in the understanding of the word ‘prediction’, because as a broader term it can be applied to any criterion situation (including concurrent). In a limited sense it may only refer to a specific prediction over a particular time interval. It is that narrow sense of the word that they use for the definition of predictive validity. In summary, the term ‘predictive validity’ should be understood as a level of fulfilment of a certain criterion after a certain time interval.

Establishing the predictive validity of a test was not a common practice; however, the examples of research in this field in the medical education literature are increasing in number. For example, predictive validity was researched for the Quebec Licensing Examination (‘QLEX’) (Tamblyn *et al.*, 1998, 2002, 2007; Wenghofer *et al.*, 2009), the United

States Medical Licensing Examination ('USMLE') (Hojat *et al.*, 2007; McCaskill *et al.*, 2007), the BioMedical Admissions Test ('BMAT') (McManus, Ferguson, Wakeford, Powis, & James, 2011), the Medical College Admissions Test ('MCAT') (Coumarbatch, Robinson, Thomas, & Bridge, 2010; Donnon, Paolucci, & Violato, 2007; Dunleavy, Kroopnick, Dowd, Searcy, & Zhao, 2013; Julian, 2005), and many others, for example the Flemish Admission Exam (Lievens & Coetsier, 2002). According to Hutchinson, Aitken, & Hayes (2002), between 1985 and 2000 there were fifty-five publications on validity in the field of medical education, of which only two related to the issue of predictive validity. In comparison, fourteen considered concurrent validity issues. An analogous review<sup>2</sup> of publications between 2000 and 2014 performed solely for the purposes of this thesis showed that among sixty-nine publications found (see Appendix B), thirty-two related to certain aspects of predictive validity. This suggests that there is an increasing interest in the predictive value of educational tests. The published research approaches the process of collecting predictive validity evidence for medical exams in two ways: via comparison with clinical outcomes, and via comparison with other medical exams.

#### **1.3.2.1 Relation between an exam and clinical outcomes**

The first group of predictive validity criteria are the results of clinical decisions, or alternatively, measures of quality of clinical treatments. There is an extensive number of such potential clinical outcome measures that can be applied to research. For example, Lindner and colleagues devised twenty-three highly clinical indexes of quality of care, of which seventeen were considered reportable (Linder, Ma, Bates, Middleton, & Stafford, 2007; Ma & Stafford, 2005). Others, like Tamblyn and colleagues (Tamblyn *et al.*, 2007), used complaints as a criterion. Based on the literature, criterion measures were categorised into three major groups:

1. Measures of malpractice e.g. complaints against medical professionals (Tamblyn *et al.*, 2007), or disciplinary actions (Papadakis *et al.*, 2005; Papadakis, Hodgson, Teherani, & Kohatsu, 2004; Papadakis, Arnold, Blank, Holmboe, & Lipner, 2008).
2. Specific measures of clinical performance e.g. morbidity rate (Kelly & Hellinger, 1986), mortality rate (Norcini, Boulet, Opalek, & Dauphinee, 2014; Norcini, Kimball, & Lipner, 2000; Norcini, Lipner, & Kimball, 2002), mammography screening rate

---

<sup>2</sup> The review was performed based on a similar methodology to the one presented by Hutchinson *et al.* (2002). The description and the list of papers with references to the area of validity or psychometric qualities constitute Appendix B.

(Pham, Schrag, Hargraves, & Bach, 2005; Tamblyn *et al.*, 1998, 2002), or completion of prenatal visits ratio and low birth weight (Haas, Orav, & Goldman, 1995).

3. Subjective ratings of clinical performance e.g. information from patient satisfaction questionnaires, perceived quality of medical services, and peer skills assessment (Ramsey *et al.*, 1989; Reid & Friedberg, 2010; Wenghofer *et al.*, 2009) in clinical contexts.

The findings from the predictive validity studies employing the above-listed criteria seem to support a notion that certification and licensure exams add to the perceived quality of medical services (Reid & Friedberg, 2010), and significantly increase the perceived competence of medical professionals (Holmboe *et al.*, 2008). The results of a meta-analysis by Sharp, Bashook, Lipsky, Horowitz, & Miller (2002) demonstrated that out of thirty-three such studies, sixteen showed significant positive association between certification and better clinical outcomes, three studies showed negative association (in the authors' opinion this was mostly due to a lack of case-mix adjustment), and fourteen showed none. The study was criticized after a secondary statistical analysis by Grosch (2006), who did not find enough evidence to sustain Sharp and colleague's conclusions that linked better care with licensing exams. In his paper Grosch was very strict on applying Hill's causal criteria (Hill, 1965) for establishing causality between examination and better clinical care, and negated the value of retrospective longitudinal studies due to the data-dredging bias (Sackett, 1979). Hence, his critique addressed both particular studies referenced by Sharp *et al.* and their selective choice, with the main focus on the latter. Grosch's comments are not without merit and should be considered, although he failed to propose a viable alternative of a robust study. In fact, meeting all of Hill's criteria may not even be possible in the field of medical education where experimentation, mentioned as one of the Hill's criteria, is not always feasible. Grosch also seems to disregard consistency and coherence among Hill's criteria, as the body of knowledge on the relationship between licensing exams and quality of care is more extensive than the limited number of studies mentioned by Sharp *et al.* For example, it was shown that doctors who scored higher in QLEX and Medical Council of Canada Qualifying Examination ('MCCQE') had higher mammography screening rates, which was in turn associated with better care (Tamblyn *et al.*, 1998, 2002). This was further supported in a study by Pham and colleagues (Pham *et al.*, 2005) who indicated that certification was linked to higher ratio of delivering preventive services to patients. Studies by Norcini, Kimball, and Lipner (Norcini *et al.*, 2000, 2002), of which one was referenced by Sharp *et al.*, suggested that lower patient mortality was associated with certified doctors in

comparison to non-certified doctors. In a recent study on performance of international medical graduates ('IMGs') Norcini and colleagues (Norcini *et al.*, 2014) found that lower scores on the American USMLE Step 2 clinical knowledge exam are associated with higher mortality rates among patients with congestive heart failure and acute myocardial infarction. In addition, Rutledge *et al.* (1996) found that patients treated for ruptured abdominal aortic aneurysm by certified surgeons had higher survival rates. A study by Tussing and Wojtowycz, also referenced by Sharp, found that certified obstetricians had a higher caesarean section rate than non-certified doctors, which was associated with higher qualifications of the doctors (Tussing & Wojtowycz, 1993). Moreover, Levy, Mohanaruban, and Smith (2011a) found a significant positive correlation between MRCP(UK) and Workplace Based Assessment outcomes ('WBAs'), which also assess a clinical component. Similarly, doctors who obtained low results in the Canadian MCCQE Part 1 and MCCQE Part 2 were statistically more likely to obtain an unacceptable outcome in a peer assessment on quality of care (Wenghofer *et al.*, 2009). This supported the findings of a study by Ramsey *et al.* (1989), who found that certified physicians were better in other written knowledge examinations. Supporting Ramsey's findings, Tamblyn and colleagues (2002) found that doctors who obtained higher scores in drug knowledge tests on MCCQE had a lower risk of contraindicative prescribing. Considering the number of publications in the field and the consistency of the findings, despite Grosch's objections, it was acknowledged that medical certification examinations are likely to be linked with better care, and therefore, clinical measures may be considered suitable criteria for a validity study.

#### **1.3.2.2 Relation between two medical exams**

The drawing of comparisons between test results and other medical exams is the second method of predictive validation employed in the examples from the literature. Nibert, Young, and Adamson (2002) found that Health Education Systems Incorporated Exit Exam predicted results in National Council Licensure Examination ('NCLEX'). Swanson, Case, Koenig, and Killian (1996) found that US MCAT scores were predictive of licensing exam results (from  $r=0.14$  in writing samples of the exam, to  $r=0.54$  in biological sciences). These conclusions were confirmed later by Donnon, Paolucci and Violato (2007) in their meta-analysis. Similarly, McManus *et al.* (2011) found that BMAT Knowledge and Applications was predictive of academic performance in medical school; however, the predictors were weak (from  $r=0.36$  in year 1 to  $r=0.23$  in year 2). Analogous results were found by Lievens and Coetsier (2002) for the Flemish Admissions Exam "Medical and Dental Studies" in cognitive ability tests (from  $r=0.11$  to  $0.13$ ). Therefore, the consensus view of the literature

is supportive of the methodology of comparison between two separate medical exams as a method of establishing predictive validity.

It might be argued that comparing results of one exam to the results of another may only indicate ability to pass exams well, as the reflection of general aptitude. However, such an approach would have to disregard the complexity and extent of medical knowledge. It would also have to disregard the results of studies where general aptitude did not predict – or only predicted to a small extent – the results of exams. In particular in a study by McManus, Woolf, Dacre, Paice, & Dewberry (2013), it was found that AH5 intelligence test results did not correlate significantly with school finals ( $r=0.05$ ), or MRCP(UK) Part 1 results ( $r=0.13$ ). The aptitude part of BMAT was found to be non-predictive of educational achievements in year 1 and 2 of medical school (McManus *et al.*, 2011), or predicted those results to a very limited extent ( $r=0.15$ ) (Emery & Bell, 2009). Similarly, in the United States the Scholastic Aptitude Tests ('SATs') were not predictive when other measures of scholastic achievement were taken into account (Baron & Norman, 1992). A meta-analytic study comparing predictive validity of the attainment and aptitude tests further showed that aptitude tests generally predict future performance less well in comparison to the educational knowledge tests (McManus, Dewberry, *et al.*, 2013). Further, it was previously found that academic achievements predicted occupational success (Barrett & Depinet, 1991). This pattern was discussed by McManus and colleagues who referred to it as the 'Academic Backbone' (McManus *et al.*, 2013). The authors of that study argue that there is a positive predictive relationship between any two examinations in the academic path, and they also suggest that previous achievements constitute a base that further knowledge is building upon; a concept referred to as 'cognitive capital'. This clearly relates to both the concept of expertise and the concept of general ability. The authors further analyse the effect of intelligence and various motivational factors on medical education achievements, indicating they should be positively correlated. However, as they argue, intelligence, personality and motivation alone are not sufficient to succeed in medicine. Based on the above, exam results can be used as criterial measure in a validity study, with certain limitations associated with the methods they are conducted with, i.e. after consideration of common method variance issues.

#### **1.3.2.3 Common method variance**

It has been argued that drawing a comparison between any two measures may be affected by common methods variance (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003); meaning that the relationships between them depend not only on the coherence between the



constructs represented by these measures, but also, among others, the form of the measurement (Cronbach, 1970), response process, judgment bias, temporal affects, and cognitive biases, such as halo effect or social desirability (Crowne & Marlowe, 1964). The similarity of the forms and the coherence of the constructs, combined with the effects of biases, may be hard to distinguish in a research process and may pose a challenge to the interpretation of statistical results. It was found by Cote and Buckley (1987) that as much as 26% of variance, which corresponds to a correlation coefficient being approximately 0.50, can be associated with systematic and random errors being the result of common method biases. Although this number might be considered worrying and could potentially have an effect on conclusions drawn from this research, there are several strategies that allow for minimising that risk.

Podsakoff *et al.* (2003) named, among other methods, assurance of anonymity, counterbalancing of items, or ensuring variability in the data collecting personnel (or in the case of MRCP(UK) or other medical exams, examiners) as ways to alleviate some of the problems. The designs of medical exams apply the majority of the techniques proposed by Podsakoff *et al.*, especially in relation to collecting data, response process, and item characteristics. In the case of this particular validity study, however, minimising the effect of common method variance on the interpretation of the results required further undertakings. In particular, two other techniques named by Podsakoff *et al.* were relevant. One was based on obtaining validation data from multiple sources, and therefore, it was a goal ahead of this research to obtain data from as many medical exams and assessments as possible. Second, this research aimed to ensure triangulation of measurement methods for obtaining validation criteria in accordance with the Multitrait Multimethod methodology proposed by Campbell and Fiske (1959). Following their approach, it is crucial to present evidence of relationships between the predictor and the criteria that measure coherent and divergent constructs obtained with multiple methods. Ensuring wide selection of measurement approaches of these constructs minimises the risk of the correlations being the result of common method variance, as discussed by Podsakoff *et al.* Hence, choosing both examinations and clinical criteria obtained from variety of sources and measured with diverse methods should minimise the risk of validity inferences being insufficiently substantiated. The details of the process of selecting the criteria is further described in Chapter 2.

## **1.4 SCOPE OF THE WORK**

The main purpose of this PhD research was to gather and evaluate the evidence for the predictive validity of MRCP(UK). This subject carries significance not only for the doctors who take MRCP(UK) or the medical profession to provide means to address the issue of its accountability, but indirectly it may be a confirmation to the general public that the system of ensuring a good standard of medical care works effectively. This research investigating the validity and quality of MRCP(UK), however, was not aimed at examining the condition of the entire UK healthcare system, nor at defining what a good doctor is. Although both of these issues are indeed important and constitute a major piece of background for this research, as briefly addressed in the introduction, the purpose of this research was much more specific. This was mainly because this thesis sought an answer to the question of whether MRCP(UK) assesses the components of what the literature considers to be the professional knowledge, skills and attitudes of a potential physician.

In order to test whether MRCP(UK) selects candidates who possess those qualities and are able to deliver high-level medical services – and de-selects those candidates who do not possess those qualities – the research should be able to examine if external criteria congruent with the three key qualities of a professional are predicted by the MRCP(UK) scores. This would constitute firm evidence that passing MRCP(UK) indeed translates into better medical care. The Royal Colleges of Physicians recognised the need to validate the MRCP(UK) results through its predictive value and funded a PhD studentship to examine the issue, which is why this research was focused solely on MRCP(UK), and not on medical exams in general. Notwithstanding this limitation, the conclusions from this research could potentially reflect on other medical assessments as well, as providing evidence of the validity of one exam at the same time provides evidence in favour of validity of those that were chosen as the criteria.

The MRCP(UK) exam is one of the largest physician examinations in the world. Due to the large number of doctors attempting MRCP(UK) annually and a long tradition of its administration, a significant amount of reliable longitudinal data were secured by the RCPs. The extent of that data allowed for extensive analyses and the drawing of certain inferences applicable to postgraduate medical educational testing in general. This constitutes yet another reason why the results of this study may add to the body of knowledge on high-stakes examination in medical education.

## 1.5 AIM AND HYPOTHESES

Although in the case of MRCP(UK) a wide array of psychometric issues have already been extensively addressed (Dacre, Besser, & White, 2003; Dewhurst, McManus, Mollon, Dacre, & Vale, 2007; Elder *et al.*, 2011; McManus & Lissauer, 2005; McManus, Elder, *et al.*, 2008; McManus, Mollon, Duke, & Vale, 2005; McManus, Mooney-Somers, Dacre, & Vale, 2003; McManus, Thompson, & Mollon, 2006; Tighe, McManus, Dewhurst, Chis, & Mucklow, 2010), so far little research on the predictive validity of the examination has been conducted (Levy *et al.*, 2011a). The purpose of this doctoral thesis was, therefore, to examine the existing evidence for the predictive value of MRCP(UK).

Similar to other large licensing or certifying medical exams such as QLEx, MCCQE or USMLE, MRCP(UK) is a high-stakes exam, with both written and practical examination components. Based on these similarities to other medical exams, in view of relevant publications and in consideration for the construct that MRCP(UK) is designed upon, it was hypothesised that:

**Performance on MRCP(UK) will predict:**

- 1. future performance in other assessments in medicine of knowledge and clinical skills,**
- 2. measures of clinical performance or underperformance**
- 3. measures of professional behaviour or lack thereof,**

**as available in the UK setting.**

In view of Campbell and Fiske's Multitrait Multimethod Matrix (Anastasi & Urbina, 1997; Campbell & Fiske, 1959) the relationships between similar constructs should be higher than between two different constructs. This means that coefficients representing the relationship between two knowledge tests or two clinical assessments should be higher than between mixed pairs of, for example, a knowledge test and a clinical assessment.

It was also predicted that doctors obtaining higher scores in MRCP(UK) exams would present a better professional standing, and would be better in subsequent knowledge exams. Established relationships were predicted to be of moderate strength, as in accordance with the psychometric theory (Cronbach, 1970, p. 137) the strength of an uncorrected validity coefficient is unlikely to exceed 0.60.

## SUMMARY

Following the assumed standards of medical practice based on the concept of the New Professionalism, all aspects of medical profession activities that include medical education have gone through changes that aimed to improve the quality of medical services and facilitate increase in public trust. Exams are considered an important part of the educational process, as they allow verification that the competence of those who take them meets the set standards of the profession. Therefore, those exams should be good measurement tools and their quality should be verified. High-stakes exams such as MRCP(UK), which is comparable to other high level medical certification exams, are particularly important due to their effect on individual lives and on society. Examination of evidence on predictive validity of MRCP(UK) can help confirm that MRCP(UK) indeed measures qualities associated with the notion of a medical professional. Predictive validity can be assessed through a variety of methods. However, the examples from the literature provide two main approaches: through comparison with clinical measures of performance, and through comparison with other medical exams. It was hypothesised that MRCP(UK) would predict both types of measures. If verified positively, the evidence of predictive validity of MRCP(UK) would imply it is a good exam, which may justify its existence and grant its continuance if challenged. With the discussion on the meaning of selection procedures and exclusivity of the medical profession, and in view of an increasing concern about the quality of the healthcare system, the confirmation of the merit of MRCP(UK) may constitute an argument that the medical profession takes necessary steps to minimise the risks to patients' health, which may lead to an increase in public trust.

## Chapter 2

### Specificity of the UK training system and description of the data sources

#### ABSTRACT

*In order to facilitate the understanding of the methods employed in this research, this chapter discusses the potential and secured sources of validity criteria originating in the UK medical exams and clinical work. First, an overview of the UK medical education system is presented, with a particular emphasis on the education of physicians. Next, a description of medical career paths is presented providing an explanation on the selection of potential sources of criterion measures. A discussion of the sources of data in light of the three components that MRCP(UK) aims to assess, namely: knowledge, clinical skills, and attitudes, is provided. Subsequently, the sources of data that were secured for the purposes of this research are described in detail. The information on those exams provides information on the role they fulfil in the medical education system, their design, psychometric features (in case of exams) or quality of the data (in case of registers and formal reviews). In particular, this chapter presents descriptions of the MRCP(UK) exam with its history, the MRCGP exam, the Clinical Oncology (FRCR) exam, the Specialty Certificate Exams and the Cardiology Knowledge Based Assessment, the Annual Review of Competence Progression process and its outcomes, the List of Registered Medical Practitioners, and the General Medical Council's Fitness to Practice procedures.*

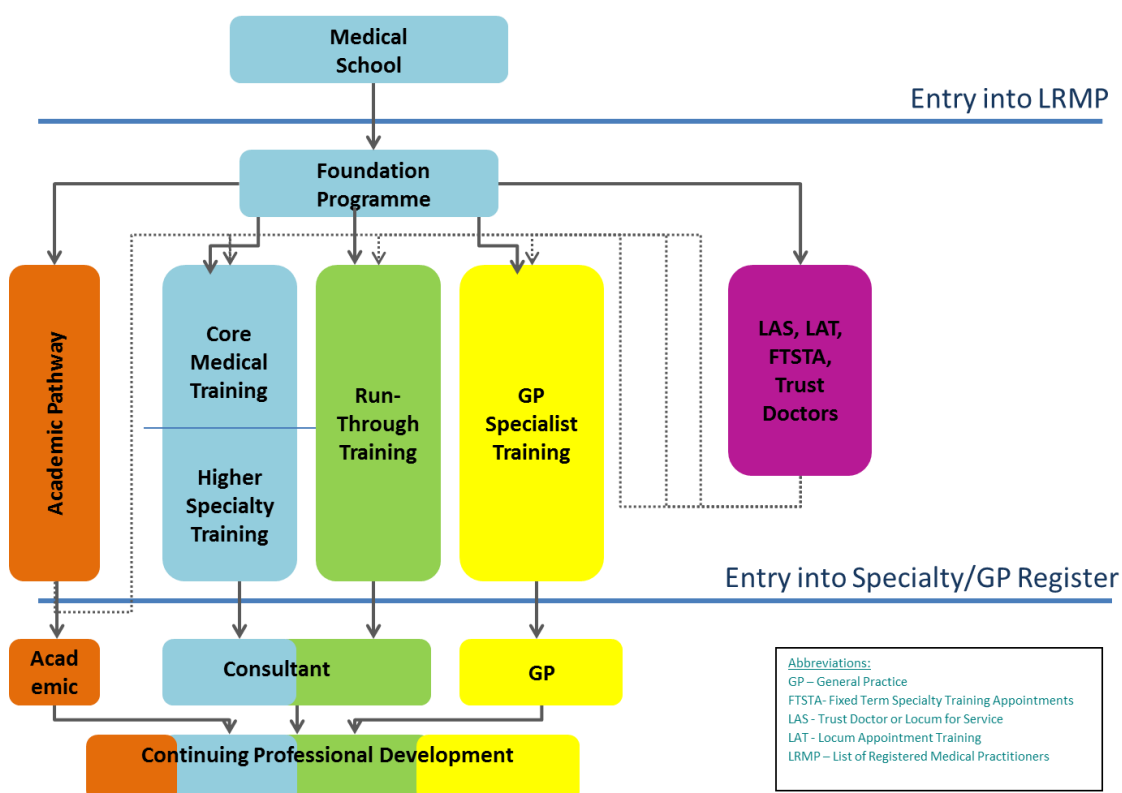
MRCP(UK) is an examination designed to select doctors with the appropriate knowledge, skills, and attitudes to become qualified physicians. In order to conduct a predictive validity study, the chosen criterion measures should relate to each of those constructs. The criteria should also comply with theoretical notions of validity (Anastasi & Urbina, 1997). The relationship between any chosen criterion and performance in MRCP(UK) should be logically congruent (Brennan *et al.*, 2004). Any chosen criterion should also be a reliable measure. Finally, the criterion measures should be UK-specific, due to the fact that MRCP(UK) is designed to fit in the UK educational system. Meeting these recommendations would support the credibility of the obtained results. Further, based on the discussed literature, the criteria chosen for this research were required to be embedded within the

medical education or health services system in the UK and should have been either clinical performance measures including professional behaviour assessments or knowledge assessments. In order to justify the choice of criteria, a short overview of UK medical education with a particular stress on physicians' educational path seemed indispensable.

## 2.1 MEDICAL CAREER OPTIONS: AN OVERVIEW OF THE MEDICAL EDUCATION

### SYSTEM AND PHYSICIAN'S TRAINING

The medical education system in the UK provides doctors with a variety of available career paths. Figure 1 presents a diagram with typical training paths.



Source: based on NHS Medical Careers(<http://www.medicalcareers.nhs.uk/>)

Figure 1. A simplified overview of typical postgraduate medical training paths in the UK.

The first stage, at the top of the diagram, is the completion of an undergraduate medical degree upon which a doctor enters the Foundation Programme for two years, and registers with the GMC on the List of Registered Medical Practitioners ('LRMP'). Doctors then enter specialty training, with several options to choose from. The colours in the figure above denote different consistent choices. The solid arrows show automatic progression between stages upon meeting training requirements. For example, orange represents the academic

career path (e.g. researchers), while purple denotes temporary career choices, such as Fixed Term Specialty Training Appointments ('FTSTA'), Locum Appointment Training ('LAT'), Trust Doctor, Locum for Service doctor ('LSA'), or teaching fellows. Within the main specialty stream, the blue rectangles represent uncoupled training, which is a path typical for physicians or surgeons, yellow rectangles represent training typical for a general practitioner, and green rectangles show run-through training, which is characteristic for such specialties like anaesthesiology or psychiatry. Temporary career choices may lead to the major specialty stream (paths marked with dotted arrows); however, doctors may at any point opt out, or may not make it to their chosen specialty.

Each career choice is bound with certain requirements of a body that supervises the specialty. Passing an entry or completion exam could be such a requirement. For example, assuming a doctor wishes to become a physician, they need to pass all parts of MRCP(UK) during the Core Medical Training ('CMT') to become a member of the college and to progress to the higher Specialty Training ('ST'). With a certificate of completion of training ('CCT') they may become a specialist consultant or occupy other senior hospital position(s). At this point, their further education requires maintaining professional standards through updating prior knowledge, which involves gaining additional skills, learning additional specialities, and undergoing revalidation procedures. This is referred to as continuing professional development ('CPD').

This complex system of career choices is organised and supervised by different entities. Focusing on their key responsibility, Figure 2 presents a general diagram of relationships between major institutions of that system in relation to the stages of medical training and type of activity<sup>3</sup>. This system was in place until March 31<sup>st</sup> 2013.

---

<sup>3</sup> This system was functioning during the data collection process and analyses.

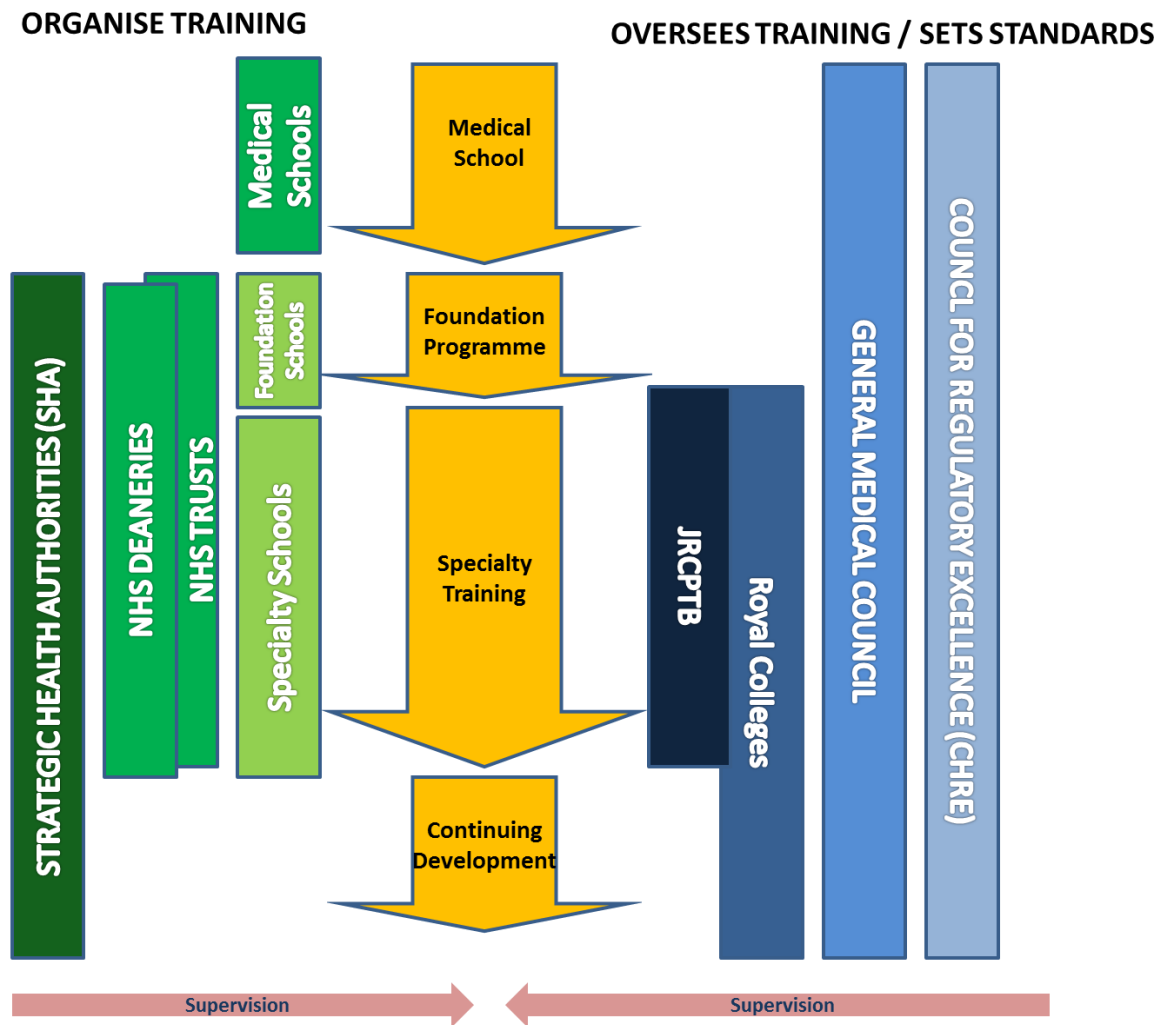


Figure 2. Division of responsibilities between major institutions influencing the training of a doctor in the United Kingdom.

The yellow arrow in the middle of the graph represents the key stages of medical education based on the uncoupled major specialty stream. On the diagram, the further the institution is located from the yellow arrow, the more broad its impact on medical training. Pink horizontal arrows indicate the chain of supervision.

The green fields on the left side of Figure 2 were the institutions responsible for the design and execution of the educational and training curricula (until March 31<sup>st</sup> 2013). The Medical Schools were responsible for the undergraduate medical training and the Primary Medical Qualification ('PMQ'). Further training was organised in the hospitals, which were under the administration of the NHS Trusts, but the training was also supervised by the NHS Deaneries. The NHS Trusts reported to Strategic Health Authorities ('SHAs'). As of 1<sup>st</sup> of April 2013 the SHAs were no longer in existence due to changes introduced by the Health



and Social Care Act of 2012. Part of SHAs' responsibilities were transferred to the NHS Commissioning Boards, but the responsibility for training supervision was transferred to NHS Health Education England and the 13 Local Education and Training Boards ('LETBs'). However, as 2013 was intended as a transition year, the division and the shape of responsibilities were not clearly defined at the moment of writing this thesis.

The institutions on the right hand side (in blue), provided guidelines for training, participated in the design of curricula, administered exams, issued licences or qualifications, and supervised the training process and the quality of medical services. The responsibilities of those institutions did not change with the new Health and Social Care Act of 2012. Therefore, for example, at the stage of CMT, where training is taking place in hospitals in the Specialty Schools under Deaneries and Trusts, the Joint Royal Colleges of Physicians Training Board ('JRCPTB') and the Royal Colleges were – and still are – responsible for supervision and examinations. A major supervisory body overseeing the whole of the training process is still the GMC, which reports to the parliamentary Council of Healthcare Regulatory Excellence ('CHRE').

Although it may appear otherwise, the GMC is entirely independent of government, and its legitimacy stems from the right of self-regulation of the medical profession. The GMC was first created in 1858, with a purpose to “protect, promote and maintain the health and safety of the public by ensuring proper standards in the practise of medicine” (GMC, 2013c). Its supervisory role is conducted through setting standards of good medical practise, providing guidelines on medical curricula and medical training, and overseeing the quality of medical services. Quality is maintained through, among other things, assessing the current level of doctors' competences through revalidation and investigating cases of practitioners whose clinical performance raised concerns. Finally, the GMC is also responsible for registering doctors, administering licences to practise medicine, and revoking any licences whenever results of clinical practice are non-satisfactory.

## **2.2 MEASURES OF PERFORMANCE IN THE UK SETTING**

There were three key issues that affected the selection process of the criteria for this research. First, this study was designed to be a purely quantitative retrospective longitudinal study devoid of qualitative components. The reasons for assuming this subjective approach are presented in section 3.9 addressing the limitations to this research. As such, the quest for criterial measures was limited to quantifiable rather than qualitative or descriptive measures.

Second, the complexity of the educational setting in the UK made the process of obtaining viable sources of data laborious. Despite the fact that all institutions from Figure 2 work closely together, the data on trainees' qualifications, all attempted exams, potential fitness to practice proceedings, peer assessments and patients' opinions, etc., were stored separately. A centralised bank containing such data is currently under design with e-portfolio (NHS, 2010b) being a start; however, at the moment of writing this thesis the data stored in e-portfolio were insufficient to be included in this study. As MRCP(UK) is usually attempted after Foundation Year 1, there were only a few institutions that could potentially provide data for the study: the GMC, the Royal Colleges, the Deaneries, and the Trusts; these were the only entities that held data on post-MRCP(UK) stages of training.

Third, MRCP(UK) was designed to test candidates for the three key aspects that define what a professional is. Those constructs are often quantified through assessments. The demands of this study were that any chosen criteria should have been the representation of similar constructs. Therefore, the criterion measures were sought among medical knowledge tests as appropriate measures of medical knowledge, assessments of clinical skills and on-the-job performance measures representing clinical ability, and assessments of behaviours that would indicate the appropriate professional attitudes.

For the reasons mentioned above, the number of potential criterion measures that could have been chosen was limited, and their selection process is described below

### **2.2.1 Assessing knowledge**

It has been widely indicated in the literature that knowledge, irrespective of its specific types (e.g. biomedical and clinical), is crucial to clinical judgment and clinical performance (Boshuizen & Schmidt, 1992; Holmboe *et al.*, 2008; West *et al.*, 2007). Knowledge as a construct can be represented by achievements in knowledge tests. A consultant physician needs to pass at least two major exams during their training: MRCP(UK), and a Specialty Certificate Exam ('SCE') or equivalent. However, that number may increase depending on the doctor's career path. Figure 3 shows educational paths in the form of a timeline, with exams that could serve as comparison criteria.

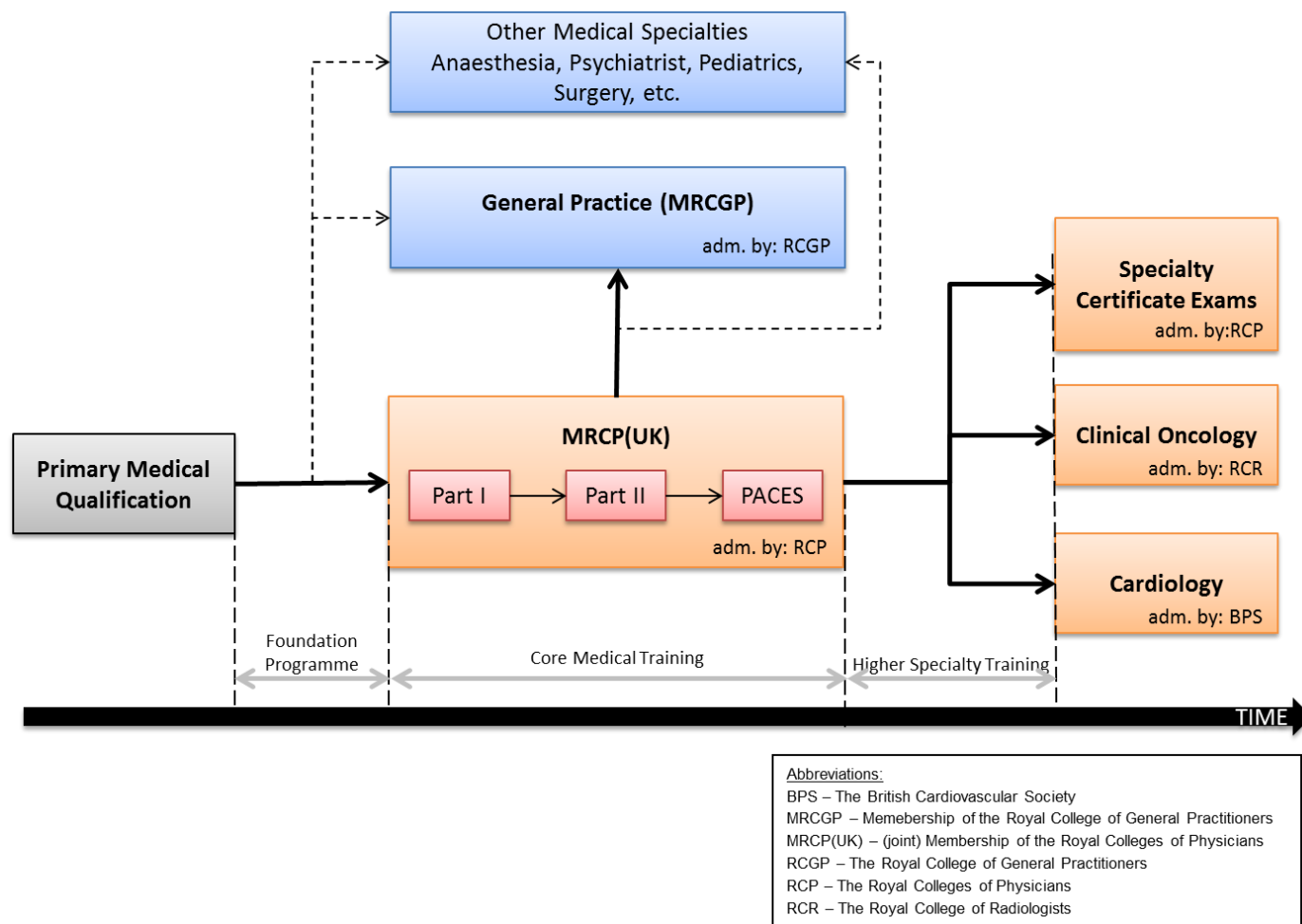


Figure 3. Description of potential postgraduate examinations that can be taken by doctors in the United Kingdom.

Orange rectangles indicate exams that are required to be taken by doctors who wish to become physicians. The blue rectangles represent exams other than those associated with the career of a physician; these exams are also administered by other colleges. The dashed arrows represent the choices that were beyond the interest of this research, or not followed-through in this research. For example, this included choosing a college other than the RCP immediately after medical school, which in fact meant that it occurred pre-MRCP(UK). This elimination was necessary in accordance with the definition of predictive validity, where evidence must be sought after a certain time-interval. The solid-line arrows are choices that were congruent with requirements of predictive validity – made post-MRCP(UK) – and followed-through in this research.

In general, doctors who take the career path of a physician in order to become consultants need to take a Specialty Certificate Exam ('SCE') or equivalent, e.g. the British Cardiovascular Society Knowledge Based Assessment ('CKBA'), or the exam for the Fellowship of the Royal College of Radiologists' in Clinical Oncology ('FRCR'). These exams are attempted approximately five to six years – assuming full-time successful training – after completing MRCP(UK), and are the final knowledge assessments in the specialist training. Access to results from the twelve SCEs, CKBA and FRCR was granted in the extent described in detail in Chapter 3.

Candidate physicians who decided to change their career path after taking MRCP(UK) did not take the SCEs. Therefore, other criterion measures representing knowledge were sought, such as another college entry exam. The abundance of colleges and medical specialties – Psychiatrist, General Practitioner, Anaesthesiologist, etc. – translated into a variety of exams that could potentially have served as criteria. All such exams were assumed to be valid sources of data, however, their use was conditional upon approval for access. The dashed arrows on Figure 3 therefore represent such paths of training where data were not accessible for the purposes of this study. Apart from the above-mentioned exams, other data were available only from the Royal College of General Practitioners ('RCGP').

### **2.2.2 Assessing clinical skills**

Skills described in *Good Medical Practice* (GMC, 2012b) include, among others, the ability to communicate effectively and politely with patients and colleagues, to make differential diagnoses, to manage cases effectively, and to take a case history. In Miller's pyramid (Miller, 1990), clinical skills represent the two top level of competence i.e. 'performance'

(shows how) and 'action' (does) (Norcini, 2003b). The ultimate level of competence means application of the obtained knowledge and skills in practice; however, the clinical skills are usually assessed on the lower levels of the pyramid, either through portfolio review, workplace based assessments, or structured clinical examinations (Metcalf, 2012). These were the potential sources of criterion measures for this research. Ideally, such measures should directly reflect proficiency in each clinical skill separately. However, clinical skills often interact and jointly contribute to general clinical performance and jointly aid proper clinical situation management. Hence, the majority of the criteria in the literature do not differentiate between certain skills. For example, misdiagnosis, contraindicative prescription, morbidity or mortality rates were previously used as valid criteria (see section 1.3.2.1 for referenced studies), even though it was impossible to distinguish which of the specific clinical skills affected these criteria the most. Therefore, following these examples, this study sought general clinical performance measures as well.

On review of the literature there appears to be a variety of such tailored indicators of clinical performance useful in a predictive validity study. However, despite this theoretical abundance, the available sources of clinical performance criteria for the study on predictive validity of MRCP(UK) were very limited. Contrary to the situation in the United States or Canada (from where most examples originate), the specificity of the UK's medical system is that institutions monitoring the quality of medical care are currently unable to calculate mortality rates, morbidity rates, contraindicative prescribing incidents, and other direct clinical outcome measures for individual physicians. Upon discussion of introducing such measures of individual performance in the UK there was a considerable amount of concern from doctors, who argued that such measures might not be properly risk-adjusted or would not describe the context of doctors' decisions. They reasoned that this then could lead to misinterpretation and misunderstandings due to lack of proper risk-adjustment or contextual factors (Lanier, Roland, Burstin, & Knottnerus, 2003). The current situation directly results from addressing these concerns. The applied quality assurance procedures in the UK do not allow for the collection of such data, which seriously limits the potential to directly assess clinical performance.

The quality assurance process, however, requires that medical services be monitored. Based on an on-going evaluation process criteria referring to clinical underperformance were identified. Also, extending from the arguments in favour of employing exams as suitable criterion measures, standardised assessments of clinical skills or training

performance were assumed to be a viable source of data for this validity study. A detailed description of both types of clinical criteria is provided in the following sections.

#### **2.2.2.1 Underperformance and the List of Registered Medical Practitioners**

Referring to the examples from the research conducted in the US (Papadakis *et al.*, 2005, 2004; Papadakis *et al.*, 2008; Ramsey *et al.*, 1989; Reid *et al.*, 2011; Tamblyn *et al.*, 2007; Wenghofer *et al.*, 2009) clinical underperformance can be quantified through the rate of medical errors, complaints, or counts of cases of misconduct. However, the UK setting differs significantly from the American.

Firstly, there are no detailed public records on prevalence of cases of medical errors that would allow for identification of an individual doctor, nor are there publicly accessible records for doctor-patient litigation cases. Data on the latter are collected by the Medical Protection Society; however, access to that information is bound by an extremely strict procedure that aims to prevent disclosure of any personal information that would allow the identification of an individual doctor. This limitation made linking potential litigation data with the MRCP(UK) data almost impossible. Even if the data were made available for the purposes of this research each case would have to be reviewed individually, as a lack of proper attention to its particulars could lead to misleading interpretations and wrong conclusions. Therefore, although such cases could potentially be a source of valuable data, their qualitative rather than quantitative character made their use beyond the assumptions of this project. In summary, direct data on medical errors were neither available nor, after consideration, considered a practical source of criteria for this study.

Secondly, the literature examples refer to complaints and misconduct. The records show that in 2011-2012 there was a significant number of written complaints filed to NHS England, which for all medical positions (including surgery) reached almost 49,300 (HSCIC, 2013b). Several previous studies on patient's complaints have shown that the key reason for complaint resulted from doctors having deficiencies in communicative, diagnostic and therapeutic skills (Kadzombe & Coals, 1992; Korsch, Gozzi, & Francis, 1968; Owen, 1991; Schwartz & Overton, 1987; Tamblyn *et al.*, 2007; Taylor, Wolfe, & Cameron, 2002; Wofford *et al.*, 2004), which are assessed, among others, during PACES. Despite the large potential of the complaint records with the NHS, no data referring exclusively to physicians were available, and their review would have to encompass a strong qualitative component, which was considered not feasible within the constraints of this project.

Another source of information on complaints were the formal referrals to the GMC. These can be made for three major reasons: doctor's misconduct, their health issues, or if their performance is deemed to be faulty or their clinical judgment seems to be impaired. Enquiries can be made by a member of the public, a member of the profession, and/or by a person acting in a public capacity. Prior to 1997, a doctor could be deemed unfit to practise due to misconduct (e.g. criminal behaviour) or health issues (e.g. blindness). After a change in the law, introduced in 1997, a doctor could also be deemed unfit to practise for poor performance, i.e. a deficit in knowledge or skills. Cases of misconduct are reviewed directly by MPTS. As previously mentioned in section 1.1.2, MPTS is funded by the GMC and reports to the GMC, but is an independent institution accountable directly to Parliament (MPTS, 2014). Decisions made by the MPTS are legally binding. The health related and performance complaints are separately first investigated by the teams of assessors from the GMC. Performance related enquiries may take a form of either a full performance investigation or competence tests. After consideration, the GMC case examiner may either close the case without sanctions, may issue a warning or agree on undertakings, or can refer the case to the MPTS for a Fitness to Practice panel ('FtP Panel'), who can take further actions on a doctor's registration with the GMC.

Registration with the GMC is mandatory to anyone who wishes to practise medicine in the UK. The main register is the List of Registered Medical Practitioners ('LRMP'). Registration is not equivalent to holding a licence to practise medicine; of approximately 270,000 names on the LRMP in September 2014, 28,000 names did not have a licence to practise (GMC, 2014a). A doctor may be registered without a licence to practise, when for example working in academia or other non-clinical jobs, in which they do not need to prescribe drugs or examine patients. Also, some doctors, particularly in their first year of the Foundation Programme may be registered provisionally with a licence, which means, for example, they can only work in approved training posts. When a GMC or FtP panel takes action on a doctor's registration, they may also act on their licence. Such sanctions include: erasure, conditions to the licence, warnings, undertakings, and suspension of the licence (in general the 'Licence Issues'). Some doctors under investigation decide to relinquish their licence in order to avoid disciplinary erasure or limitations of their right to practise; however, voluntary erasure may be also taken for other reasons. A separate category of grounds for erasure is failing to fulfil administrative obligations, such as paying the fees or not responding to the GMC correspondence, which falls under a different LRMP status – erasure for administrative reasons.

All such licence limitation cases and other changes in a doctor's status are eventually reflected on the LRMP. However, while a doctor is still in the process of a review only certain limitations are recorded. The majority of the status changes occur only after the GMC investigation or FtP Panel's decision (jointly 'GMC FtP'). Therefore, the GMC also maintains its own internal database where all investigated cases are recorded and described. That list of cases under investigation could potentially provide additional information from when licence limitations were not yet reflected on the LRMP. Therefore, two potential sources of quantifiable data indicating doctors' clinical underperformance could have been used in the current study: the registration status on the LRMP, and the fact of being investigated by the GMC FtP panel. However, it is acknowledged that quantification of the changes in the registration status or the sole fact of being investigated by the GMC does not take into account the context in which these processes occur and deprives them of their qualitative value. This may be particularly important for measures that are not immediately related to underperformance, such as for example voluntary erasures.

#### **2.2.2.2 Structured Clinical Exams, Training and Workplace Based Assessments**

The standardised clinical skills assessments usually constitute a part of the admission exams to the Royal Colleges; for example, the MRCGP exam includes Clinical Skills Assessment ('CSA'), while the Final FRCR exam ('FRCR2') has a clinical component. Such assessments constitute either an entirely separate mark (as in the case of MRCGP) or a partial score, which later contributes to an overall mark (as in the case of FRCR2). With the assumptions that this research is a quantitative study, clinical skills assessment were considered desirable criterion measures.

Apart from entry or final exams, throughout their training all doctors are assessed several times with Workplace Based Assessments ('WBAs'), which aim to verify the level of a doctor's clinical skills and evaluate appropriateness of their attitudes. WBAs are a longitudinal process based on repetitive assessment of medical trainees with six tools: Team Assessment of Behaviour ('TAB'), Logbook of procedural skills, Direct Observation of Procedural Skills, Mini Clinical Evaluation Exercise, Case-Based Discussion, and Developing the Clinical Teacher Assessment form. As such, WBAs are largely qualitative and formative in character and, therefore, of limited use for the purposes of this research for the reasons already provided above. Based on WBAs, completed exams, as well as other available evidence, an overall decision on the performance of a trainee throughout specialty training is made, and is called the Annual Review of Competence Progression ('ARCP'). ARCP is an



annual review that aims to assess qualitatively if a trainee's progress follows the assumed training plan. ARCP employs a standardised outcome scale which has a formative purpose (NHS, 2010a); however, it can be used as quantitative data with a certain level of caution. A more detailed description of how the outcomes were quantified is provided in section 3.3.1.5 devoted solely to that assessment.

Results from the sequential WBA assessments were not made available, but the access to the CSA results, FRCR2 results, and the overall ARCP outcomes was provided by the RCGP, the RCR and the JRCPTB, respectively, and they were considered valuable sources of data for this predictive validity study.

### **2.2.3 Assessing professional attitudes**

In accordance with the literature, professional attitude is described as “a predisposition, feeling, emotion, or thought that upholds the ideals of a profession and serves as the basis for professional behaviour” (Hammer, 2000). Following the American Board of Internal Medicine, Hammer (2000) identified such attitudes or their attributes as: altruism, accountability, excellence, duty, honour and integrity, and respect for others. Batenburg and colleagues (Batenburg, Smal, Lodder, & Melker, 1999) focused on patient-centeredness as the key concept of professional attitude. Lynch, Surdyk, and Eiser (2004) performed a systematic review of literature on professional attitudes, and apart from looking into the above mentioned aspects of professional attitudes they also distinguished a category of personal characteristics, which included emotional intelligence, personal values, empathy, and ‘other’.

Based on the above, it can be assumed that professional attitudes involve two main groups of behaviours: those exhibited to patients and colleagues, and those related to personal characteristics and values (Hodges *et al.*, 2011). Both groups of attitudes are difficult constructs to measure in the clinical context. This claim is supported by a systematic review by Jha, Bekker, Duffy, and Roberts (2007), who focused on ninety-seven papers and forty-four measures with reported validity and reliability, mainly designed to address specific aspects of professional attitudes. They found that there is little evidence that measures employed to assess professional attitudes as a whole are effective. However, they indicated several good questionnaires relating to professionalism, among which were for example ‘Attitudes towards Social Issues in Medicine Scale’ and ‘Medical Skills Questionnaire’, both addressing professionalism holistically, ‘Professional Decisions and Values Test’ and ‘The Ethics and Health Care Survey Instrument’, focusing on ethics, ‘Cynicism in Medicine

Questionnaire' and 'Cook-Medley Hostility Scale', which addressed personal values, and several other tools. Further, Kelly, O'Flynn, McLachlan, and Sawdon's study (2012) also showed that conscientiousness can be successfully measured. Nonetheless, upon review of the available sources of quantifiable data suitable criterion measures were hard to identify.

#### **2.2.3.1 Attitudes towards patients and colleagues**

Attitudes towards patients and colleagues often appear in patients' complaints (Hunt & Glucksman, 1991; Kadzombe & Coals, 1992; Lau, 2000; Wofford *et al.*, 2004), which could substantiate the use of complaints as a criterion. However, complaints were not considered a feasible source of data, as has been discussed already in the clinical underperformance section above (section 2.2.2.1). Also, complaints create too much ambiguity in their interpretation to differentiate between the actual lack of clinical skills and unprofessional attitudes. It has also been shown that patient satisfaction with medical services may be a derivative of doctor's communication skills (Buller & Buller, 1987; Korsch *et al.*, 1968; Schwartz & Overton, 1987; Taylor *et al.*, 2002; Wofford *et al.*, 2004), and not necessarily an indication of an improper attitude. Also, in certain types of medical treatments it has been shown that patients' dissatisfaction increases with prolonged hospitalisation or intensity of pain (Bourne, Chesworth, Davis, Mahomed, & Charron, 2010), which is not attitude related, but may affect a holistic assessment of a doctor by the patient. Finally, the complaints data prove difficult to analyse in the UK setting, as explained before in the clinical underperformance section. For all the above reasons, employing detailed attitude criteria based on complaints did not seem to be a feasible approach. Instead, it was assumed that GMC FtP procedures, which were already employed as a criterion for clinical underperformance would relate to both clinical underperformance and attitude equally. This decision was made mainly because the GMC FtP procedures may occur not only in the case of misconduct or a lack of knowledge and skills, but also in the case of persistent failure in communication or maintaining trust (GMC, 2014b).

Attitudes are often also assessed as a part of the general performance at work or in training, for example via TAB, which is a part of the WBA process. However, the lack of feasibility of employing WBAs and general clinical performance measures was addressed in the section above. Communication skills and the level of professional approach to patients are often specifically assessed during high-stakes medical examinations as a part of the standardised clinical assessment, and these were also already employed as measures of clinical performance (see section 2.2.2.2). For that reason it was assumed that attitudes in clinical situations and clinical ability are inseparable as constructs at the level of

measurement in clinical skills assessments. Therefore, all measures associated with clinical performance were at the same time assumed to represent attitudes, and analyses presented in this thesis do not separate these constructs.

#### **2.2.3.2 *Personal characteristics***

The second group of attitudes involves the personal characteristics of a doctor. This largely relates to those behaviours that support professional development such as empathy, integrity, the need for self-development, or conscientiousness (Chaytor, Spence, Armstrong, & McLachlan, 2012; Finn *et al.*, 2009; McLachlan, 2010; Stern, Frohna, & Gruppen, 2005). Such behaviours are also difficult to separate from an overall assessment of a doctor. Although personal characteristics have been shown to predict academic success (Doherty & Nugent, 2011; Enns, Cox, Sareen, & Freeman, 2001; Ferguson, James, & Madeley, 2002; Willoughby, Gammon, & Jonas, 1979), clinical performance (Haight, Chibnall, Schindler, & Slavin, 2012) and occurrences of subsequent disciplinary actions (Papadakis *et al.*, 2004; Papadakis *et al.*, 2008), personality traits, professional behaviours, and attitudes are not widely measured at any stage of medical training in the UK. Even when observed in high work ethics or meticulous approach to duties, they are usually translated into good performance in general. In a sense, personal characteristics together with motivation, can be treated as a foundation for the process of acquiring knowledge, and as an aid in clinical situations. Therefore, identifying potential criterion measures for this group of attitudes presented a particular obstacle. However, several studies (Kelly *et al.*, 2012; McLachlan, Finn, & Macnaughton, 2009) did employ a special index, the Conscientiousness Index, which is a standardised tool based on fulfilling administrative duties in a timely manner, such as delivering immunisation documentation and criminal records as required by the school, attendance, delivering summative feedback, etc. (McLachlan *et al.*, 2009). Similarly, all doctors in the UK need to fulfil certain duties required from them with respect to the GMC registration, such as paying their registration fees (GMC, 2014c). As a result of failing to meet the GMC regulations in that aspect, the GMC may strike a person off the LRMP for administrative reasons. Although the GMC may make that decision on other grounds than just failing to pay fees or submit paperwork, these two reasons are referred to most often. Therefore, administrative erasure was considered a proxy measure of conscientiousness; however, with a certain level of caution. It was also the only separate criterion for a personal characteristic that was identified in the process of this research. In terms of other aspects of attitudes and personality it was assumed that they were not measured directly, and that the criteria of an overall clinical performance

and clinical skills referred to the component of personal characteristics as well, which is reflected in the approach to the analyses.

## 2.3 SUMMARY OF THE SOURCES OF CRITERION MEASURES

As a summary of the UK medical education setting in the context of this predictive validity study, Figure 4 shows the diagram of potential and obtained data sources.

The arrows still represent training and career choices, as in Figure 3. The black dashed arrows represent career choices that were not tracked in the course of this research; the solid lines represent the tracked career choices, which occurred after attempting MRCP(UK). Green rectangles represent data sources to which access was granted, while grey rectangles represent data sources not included in this research. Based on the secured data sources the predictive validity of MRCP(UK) was assessed in two large blocks:

1. **the knowledge aspect of competence** was verified based on analysis of the relationships between MRCP(UK) scores and results of: the Specialty Certificate Exams ('SCEs'), the Cardiology Knowledge Based Assessment ('CKBA'), the Clinical Oncology First Exam ('FRCR1') and Final exam written and oral components ('FRCR2'), and the MRCGP Applied Knowledge Test ('AKT').
2. **the clinical and behavioural (attitudes) aspect of competence** was verified based on analysis of the relationships between MRCP(UK) scores and the results of MRCGP Clinical Skills Assessment ('CSA') and the FRCR2 clinical component, and additionally also based on the relationship between MRCP(UK) and the GMC FtP list, the ARCP rank outcomes, and the registration information from the LRMP.

A detailed description of each of these sources of data is provided in the following sections of this chapter.

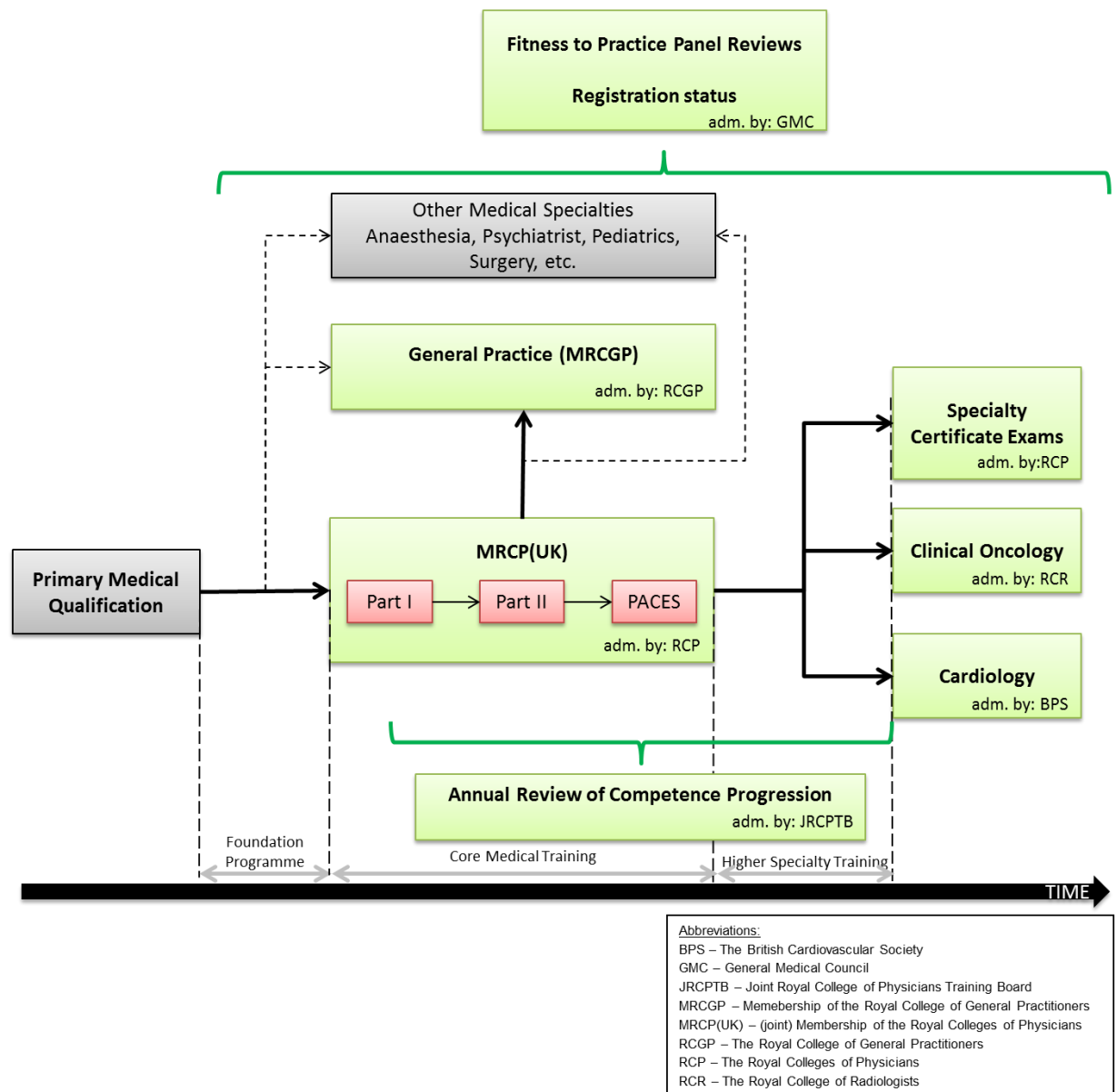


Figure 4. Possible and secured sources of criteria for the predictive validity study of MRCP(UK).

## **2.4 DESCRIPTION OF THE SOURCES OF DATA**

The chosen sources of criterion measures included medical knowledge exams, standardised clinical skills assessments, a measure of progress in medical training, and two measures of the overall clinical performance. However, in order to facilitate further understanding of how those sources of data supplied suitable measures that were applied in this study, their detailed descriptions were required. The descriptions include the purpose of the employed exams or procedures, their form, and a psychometric evaluation when available and necessary. In particular, a detailed description of MRCP(UK) is provided with a brief historical overview.

### **2.4.1 MRCP(UK)**

#### ***2.4.1.1 History of the exam***

The examination for the Membership of the Royal Colleges of Physicians of the United Kingdom is a major step in the career of a physician. The exam has a long history which has been extensively described in the literature (Clark, 1965; Cooke, 1972; Fleming, Manderson, Matthews, Sanderson, & Stokes, 1974; RCP & Cuthbertson, 2008; Waddington, 1973). The MRCP was first established after the Medical Act of 1858 to acknowledge physicians right to practice. Before the Act, the memberships of the Royal Colleges of London, Glasgow and Edinburgh served as honorary titles, which granted the right of practise within the area of influence of a particular college. Due to the exclusivity of the membership the number of Fellows was low, which guaranteed their income and status, but also limited the access to medical services. This exclusivity was partially the result of the low availability of higher education at the time, and partially due to stringent admissions procedures.

As an entry exam MRCP(UK) underwent several changes in its long history. Initially each college had their own examination and admission rules. The London college admitted Fellows and Licentiates on the basis of recommendation, university diplomas or extraordinary practice, followed by a verbal exam and a vote among the Fellows (Cooke, 1972). There are no records on the contents of the exam at that time, except that it was based on the medical knowledge of a candidate. As the admittance was voted on, no standard assessment was practiced. The Glasgow and Edinburgh colleges adopted their own independent, but similar, rules for admission. In 1771 the London College started to admit membership based on an exam in medical subjects and Greek philosophers, such as Hippocrates, Galen, and Aretaeus. The performance during the exam was still assessed with voting (Cooke, 1972).

Further modifications to the exam's form were sparked by changes in economics and society. Over the course of British history, medicine was an exclusive profession; the numbers of Fellows and Licentiates in the colleges were very limited. According to historical data the number of Fellows of the London college in the sixteenth century reached a maximum number of thirty-one, with no information on the number of Licentiates (Clark, 1965). The next available record of 1708 showed a slight increase; there were fifty-seven Fellows and thirty-nine Licentiates. In 1746 the numbers decreased again and reached fifty-four Fellows and twenty-four Licentiates (RCP & Cuthbertson, 2008).

In the nineteenth century major cities were rapidly growing in population due to industrialisation and mechanisation of production. Therefore, the number of medical professionals was insufficient for the growing needs of the working society and so the changes in the education system became inevitable. A huge pressure was imposed on changes in the medical profession admission rules and on the reform of the medical colleges (Waddington, 1973). Internal attempts to reform were ineffective and the real changes were achieved through legislation. The Act of 1858 imposed new rules of certification of medical professionals, which extensively affected the medical colleges. The exams for the membership of the Colleges became formal qualifications for practicing hospital medicine. The first new exam for the membership of the London College took place in April 1859. It consisted of four parts: three written essays of theoretical knowledge in physiology, pathology, and therapeutics, each with Latin and Greek translation passage, and an oral exam in "use and practice of medicine" (RCP & Cuthbertson, 2008). The exams were supervised by censors, who were appointed from the experienced and respected members of the college. The exam was administered in this form for the following seventy years with only slight changes related to its scope. For example, in 1867 new fields such as surgery, midwifery and the diseases of women and children, anatomy, physiology, chemistry, *materia medica*, and practical pharmacy were introduced. An example of the exam's contents from around that time is presented in Cooke (1972), and is the earliest recorded MRCP exam paper.

The exams for the Glasgow and Edinburgh Colleges were entirely autonomous. The exam for the Edinburgh College was first set in 1881 and in its original form took three days. It consisted of two written papers, a practical exam in use of medical equipment, a clinical exam, and an oral exam. The written papers addressed the knowledge of medical practise, including therapeutics, and the knowledge of "one or more departments of medicine specially professed" (Fleming *et al.*, 1974). Such specially professed fields of study included:

general pathology and morbid anatomy, medical jurisprudence, public health, midwifery and diseases of women, tropical medicine, and children's diseases. The latter two were added in 1904 (Fleming *et al.*, 1974). The clinical exam consisted of a "long case" and "short cases" (similar to London clinical exams) (Fleming *et al.*, 1974). An oral examination could address any subject from the field of medicine. Until 1886 the Glasgow College had not required any formal exam and would have admitted any doctor who could provide evidence of qualification from any recognised British or foreign university, under condition that such a candidate also obtained two thirds of the votes of the present Fellows of the College. In 1886 an exam was introduced comprising a written part in systematic medicine including pathology and therapeutics, and a clinical and oral exam (Fleming *et al.*, 1974).

Despite there being an extensive literature on expectations of the colleges towards the candidates (i.e. GMC, 1879; 1880) and on the form of exams, until 1893 the literature paid very little attention to the marking procedures. The first note was a critique of the volatile standards of marking and the leniency of the examiners, which was made by a retiring censor Dr. William Dickinson, who also suggested a unified system of marking (Cooke, 1972). Following the critique the London College introduced a standardised marking system in June 1894.

The dawn of the twentieth century brought technological and scientific progress through, amongst others, the invention of X-ray, the first vaccines, the discovery of blood types and hormones. However, the new discoveries had little effect on the colleges' admission procedures. Changes were introduced gradually. For example, between 1916 and 1924 the London College changed its rules concerning translation from foreign languages (RCP & Cuthbertson, 2008). The initial compulsory requirement of translation to Greek and Latin changed into a compulsory requirement to prove knowledge of one of four languages (Greek, Latin, French, German). The language tests became voluntary around 1925, and the points for the language assessment became a separate mark. In 1968 the language exam was entirely abandoned.

More changes were introduced between the First and Second World War. At that time the passing rates were quite low and the demand for doctors was continuously increasing. Based on historical records, in 1933 only eight candidates of seventy-five successfully passed MRCP; in 1936 the rate was twenty-five out of a hundred-and-thirty (Cooke, 1972). According to the reports of the London College, in a typical exam only 25% of the candidates were successful. This led to the next modification in the exam form to simplify



the examination process. In 1937 the MRCP started to comprise two written papers (four questions in each) and two oral exams. However, the passing rates for MRCP continued to be a huge public concern, widely expressed in the letters to the British Medical Journal. In response Robert J. Kernohan wrote (Kernohan, 1962):

It is immaterial that only 10% pass. An initial screening test might be devised to eliminate those candidates whose chance of success is negligible. Otherwise it is difficult to visualize any modification of the examination without a lowering of the high standard, which has helped to maintain the excellence of clinical medicine in British hospitals.

Following this argument, to appease the public and to address the issue of pass rates, in 1963 the London College admissions exam was divided into two parts with the first being the pre-selection exam. Shortly after, the Edinburgh College introduced a similar modification. At the time both parts comprised a written and an oral assessment. For less common specialties such as Paediatrics or Psychiatry, Part III was introduced. However, this was only offered between 1950 and the early 1970s, when the specialty colleges were established and started to administer their own exams.

In 1963 a multiple-choice computer-scored test replaced essays for the Part I written test of the London College exam. Between 1984 and 2001 that test consisted of sixty questions with five answers to be marked true or false (three hundred marked questions altogether). The questions were written by censors and in 1966 a question bank was created (Fleming *et al.*, 1974). The Part I oral exam was discontinued. Part I lasted in this form until 2002.

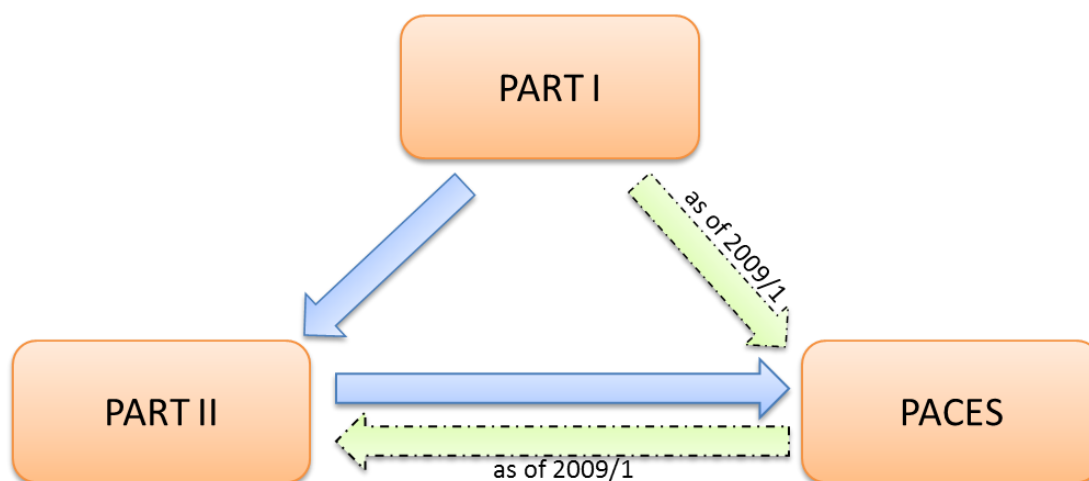
Part II changed in 1969. Essays, considered a time-consuming and unreliable method of assessment due to examiners variability, were replaced with hand-marked short open questions in three papers: case histories, data interpretation, and photographic materials (RCP & Cuthbertson, 2008). The answer key was agreed beforehand by the examiners, and the scoring method was based on the probability of answers. Examples of questions for Part II are provided in Fleming *et al.* (1974). The oral exam was kept as a middle step before allowing candidates to take the clinical exam, which also changed. The traditional “long cases” were replaced with a larger number of cases of more common illnesses. The reasoning behind the change was that the long cases were usually patients with abnormal physical symptoms that were chosen to complicate the differential diagnosis. With the limited availability of such cases for an examination candidates knew patients and cases

very well, which skewed the exam results. Replacing long cases with a larger number of patients with no abnormal physical signs for typical diseases introduced more variability to the examination tasks.

In order to provide one exam of the same standards to all medical graduates in the UK, in 1969 the MRCP exam of the London College became the joint qualification of the three royal colleges under MRCP(UK). Further changes followed. In the early 1970s negative marking was introduced in order to discourage guessing. Part II still consisted of written and clinical exams, but the written part became a multiple choice question ('MCQ') exam with a clinical focus. It tested skills in diagnosis, investigation, management, and prognosis of patients. Part II remained in this form until 2001, when the clinical assessment was separated entirely and redesigned into the Practical Assessment of Clinical Examination Skills (PACES) exam.

#### **2.4.1.2 The current form and psychometric properties of MRCP(UK)**

In its current form MRCP(UK) consists of three parts: Part I written exam ('Part I'), Part II written exam ('Part II'), and Part II Clinical Examination ('PACES'), each with three sittings (or diets) a year. Up until 2009 diet 1, Part I, Part II and PACES had to be taken and passed in the given order; however, current rules allow candidates to choose whether they wish to take Part II or PACES after Part I (see Figure 5).



*Figure 5. Explanation of the order in which MRCP(UK) needed to be attempted.*

##### **2.4.1.2.2 Part I**

Part I aims to “assess candidate's knowledge and understanding of the clinical sciences relevant to medical practice and of common or important disorders to a level appropriate for entry to specialist training” (RCP, 2013d). It consists of two three-hour papers, each

containing a hundred questions in a Single Best Answer format<sup>4</sup>. Questions cover fourteen areas of General Medicine; proportion of questions in each area is presented in Table 2.

**Table 2. Composition of Part I papers with respect to the fields of medicine.**

<i>Field of medicine</i>	<i>Number of questions overall</i>
Cardiology	15
Clinical pharmacology, therapeutics and toxicology	20
Clinical sciences, including:	25
Cell, molecular and membrane biology	2
Clinical anatomy	3
Clinical biochemistry and metabolism	4
Clinical physiology	4
Genetics	3
Immunology	4
Statistics, epidemiology and evidence-based medicine	5
Dermatology	8
Endocrinology	15
Gastroenterology	15
Infectious Diseases	15
Neurology	15
Nephrology	15
Ophthalmology	4
Psychiatry	8
Respiratory medicine	15
Rheumatology	15
<b>Total:</b>	<b>200</b>

Source: The Royal College of Physicians (RCP, 2013c)

Papers for each diet are set from the items from the question bank. Each question can be used only five times in a row; afterwards it is put aside for at least three years. The Exam Board comprising eighteen chosen members of the RCPs reviews the questions for each sitting, and assesses them based on wording, contents, difficulty, discriminative power (point-biserial coefficient) and quality of the distractor answers. Upon reaching an

<sup>4</sup>Single Best Answer ('SBA') question is a form of a Multiple Choice Question with only one answer correct. In the case of MRCP(UK) candidates are required to choose one of five choices provided. The wrong answers in SBA questions are referred to as distractors.

agreement the board members approve the contents of the paper. Papers for each diet contain approximately twenty four to thirty anchor questions; these are questions that appeared in previous diets, and which are used for statistical equating of the scores between the exams based on Item Response Theory ('IRT'). The method is statistically complex, but it is well explained by e.g. Skaggs & Lissitz (1986) or Moses and colleagues (Moses, Deng, & Zhang, 2010). Until late 2008 the passing score used to be calculated using the Hofstee method (Hofstee, Berge, & Hendriks, 1998; Norcini, 2003a), but it was changed to statistical equating, which allows for a more accurate comparison between the cohorts (RCP, 2014a).

### *Reliability of Part I*

In terms of psychometric evaluation Part I is a highly reliable exam. Between 1984 and 2001 the average reliability was as high as 0.87 (McManus *et al.*, 2003), and increased over the next years to reach 0.92.

### *Validity of Part I*

As mentioned in the introductory Chapter 1, the psychometric quality of MRCP(UK) has been extensively researched. However, published articles on validity have only referred to its specific aspects.

In terms of validity of Part I the research has shown that candidates who passed Part I had higher scores in Case-Based Discussion assessment during specialty training WBAs (Levy *et al.*, 2011a), confirming the predictive validity of Part I. Other studies found that there were sex differences in performance in the exam; however, those differences did not maintain a stable pattern. In a study by McManus and colleagues (McManus, Elder, *et al.*, 2008), in a sample of candidates attempting Part I between 1989 to 2005, the male candidates achieved higher scores than female candidates, while in a study by Dewhurst and colleagues (Dewhurst *et al.*, 2007) there was no gender related difference in performance in examinations 2003 to 2004. The changes in the observed pattern suggest that the exam is not biased against either sex. Performance in Part I was also found to be dependent on the medical school attended by candidates (McManus, Elder, *et al.*, 2008). Graduates of Oxford, Cambridge, and Newcastle-upon-Tyne medical schools were significantly better than average in Part I. Although approximately 60% of the variance between the schools was explained by pre-admission differences, when controlling for that factor, medical school still predicted MRCP(UK) scores. The studies by McManus and colleagues also indicated that performance in Part I varied depending on the cohort (McManus *et al.*,

2005). The standard of candidates increased between 1985 and 1996, and was followed by a decrease in 1997. There was a substantial difference in performance between the cohorts of 1996 and 2001. This difference was not associated with changes in the examination setting, differences in the mix of international and UK graduates or time between PMQ and the moment of taking the examination. The authors did not provide a clear explanation as to what the reasons for the decline were. However, the results supported the use of methods of setting the pass-mark other than pure norm referencing, as it was indicated that norm-referencing might have been responsible for admitting candidates of lower ability than required.

#### **2.4.1.2.3 Part II**

Part II is a written exam aimed at assessing clinical application of knowledge, e.g. case management, prognosis, investigation of results (RCP, 2013a). Before the last diet of 2005 it comprised two papers of a hundred questions each, and as of that diet it consists of three papers taken on two consecutive days. Each paper contains ninety questions in a Single Best Answer format. The questions usually include a clinical scenario or results of a medical investigation, and they might be illustrated with photographs. Example questions are available on the Royal College of Physicians website (RCP, 2013b).

Similarly to Part I, the MRCP(UK) Exam Board chooses the questions to be included in each paper based on their psychometric qualities. The questions from different fields of medicine are balanced to meet the proportions presented in Table 3.

The use of questions across diets follows the same rules as for Part I. Anchor questions are used to allow for statistical equating. The pass-mark is currently set up based on statistical equating instead of the Hofstee method. This changed at the beginning of 2010 to ensure better comparability between results of different cohorts of candidates.

**Table 3. Composition of Part II paper with respect to the fields of medicine**

<i>Field of medicine</i>	<i>Number of questions overall</i>
Cardiology	25
Dermatology	13
Endocrinology and metabolic medicine	25
Gastroenterology	25
Geriatric Medicine	10
Haematology	13
Infectious diseases and GUM	25
Neurology	25
Nephrology	25
Oncology and palliative medicine	13
Ophthalmology	4
Psychiatry	4
Respiratory medicine	25
Rheumatology	13
Therapeutics and toxicology	25
<b>Total:</b>	<b>270</b>

Source: The Royal College of Physicians (RCP, 2014b)

### *Reliability of Part II*

Part II has a lower reliability than Part I, but at an acceptable level for a high-stakes examination. Between 2002 and 2005 Part II contained a maximum of two-hundred questions and Cronbach's Alpha varied between 0.73 to 0.83. It was argued by Tighe and colleagues (Tighe *et al.*, 2010) that despite the lower reliability Part II was a reliable exam. They indicated that low Alpha for Part II resulted directly from limited variance of scores in Part II, as a result of employing Part I as a pre-selection exam. Introducing a 'sieve' exam, such as Part I, reduces the variation in the pool of candidates for the subsequent exam in comparison to the overall population of possible candidates. The formula for Cronbach's Alpha coefficient (Anastasi & Urbina, 1997) is based on candidate scores variance, which in the case of restriction of range reduces its size (Sackett, Laczo, & Arvey, 2002). Hence, the lower the variance in test scores, the lower the Cronbach's Alpha. Based on the formula (Anastasi & Urbina, 1997; Tighe *et al.*, 2010), the Standard Error of Measurement ('SEM') depends on the reliability of an exam, but also on the standard deviation of scores. The general perception is that reliability needs to be high and SEM small, and this stems directly from the equation for SEM. But since SEM depends on standard deviation, the smaller the

standard deviation is, the smaller SEM is, even though the reliability may also fall. Therefore, SEM seems to be a better measure of accuracy of results than reliability, as argued by Tighe *et al.*, as it is less susceptible to the extent of variance. Yet, SEM is not an officially recognised measure of accuracy of an exam, and despite this evidence, in order to increase the reliability of Part II the number of questions was increased based on Spearman-Brown formula (Anastasi & Urbina, 1997) in 2005. This allowed for an increase in Part II reliability coefficient to a stable level of 0.85.

### *Validity of Part II*

Comparable to the evidence for Part I, it was found that candidates who passed Part II obtained higher scores in Case-Based Discussions during specialty training (Levy *et al.*, 2011a). In terms of sex differences in performance, according to recent studies (Dewhurst *et al.*, 2007; McManus, Elder, *et al.*, 2008) there were none. Differences resulting from the medical school affiliation indicated an analogous pattern as in the case of Part I.

#### **2.4.1.2.4 Practical Assessment of Clinical Examination Skills (PACES)**

The PACES examination aims to ensure that candidates:

...demonstrate clinical skills of history taking, communicate clinical information to colleagues, patients or relatives of patients, examine patients appropriately and interpret physical signs, make appropriate diagnoses, develop and discuss emergency, immediate and long-term management plans and discuss ethical issues (Dacre *et al.*, 2003),

which is in accordance with the guidelines of *Good Medical Practice* (GMC, 2013a).

As it was mentioned previously, in June 2001 the clinical assessment within the Part II examination became PACES. It no longer contained neither long nor short cases, but instead started following an OSCE<sup>5</sup> model. PACES (oPACES; original PACES) originally comprised five twenty-minute standardized stations with real patients or surrogate patients (actors) with a variety of conditions from the following areas: Abdominal and Respiratory Systems (two cases; ten minutes each); Cardiovascular and Nervous Systems (two cases; ten minutes each); Skin, Locomotor, Eyes and Endocrine (four cases; five minutes each); History Taking; and Communication Skills and Ethics. Each station employed two examiners

---

<sup>5</sup> OSCE stands for Objective Structured Clinical Examination

who assessed candidate performance in that station based on a four-point scale: Clear Pass/Pass/Fail/Clear Fail. The minimum for passing oPACES was a total of forty-one points (Dacre *et al.*, 2003). In 2009 the assessment of the overall performance in a station was changed to an assessment of generic skills evaluated across stations (e.g. managing patient's concerns or communication) in order to prevent the occurrence of a compensatory effect between different stations, e.g. better communication skills compensating for lesser skills in physical examinations (Elder *et al.*, 2011). With that change the scale for assessment was also shortened to a three-level scale: Satisfactory/Borderline/Unsatisfactory.

New PACES still lasts a hundred and twenty-five minutes and consists of five stations of twenty minutes each, with five minute breaks in between (see Figure 6); however, Skin, Locomotor, Eyes and Endocrine station was changed in 2009 into Brief Clinical Consultation (two cases, ten minutes each). Candidates start at a random station and rotate clockwise until they have completed all five stations. Each station requires preparation: certain stations provide a written scenario as an introduction, e.g. for history taking or communication station. An example scenario is presented in Appendix C.

During the exam, PACES candidates either examine real patients or communicate with surrogate patients (actors), for example, in Communication and Ethics and History Taking stations. Each station is designed to test generic clinical skills in different areas of medical expertise.



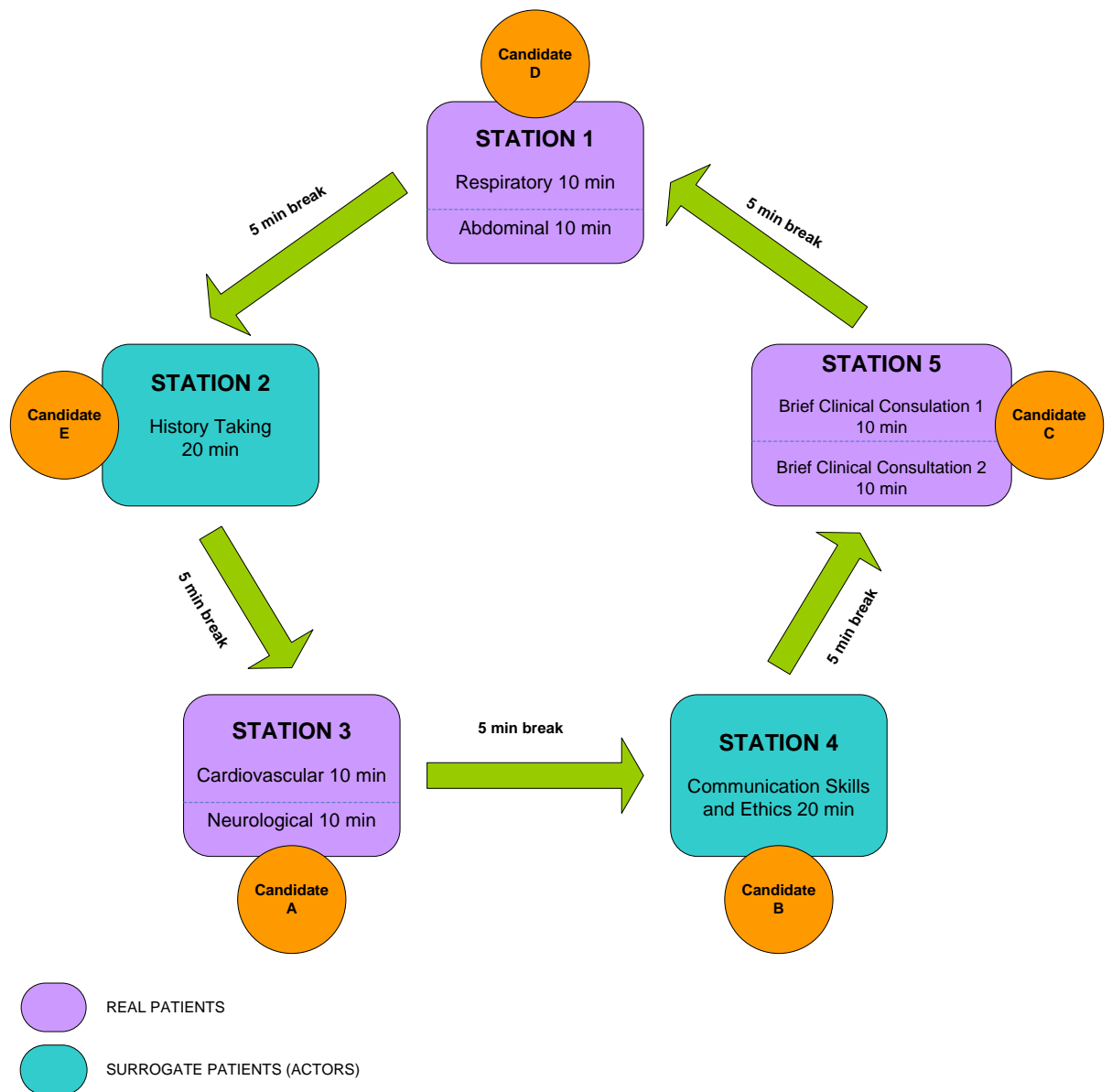


Figure 6. Diagram of current PACES station descriptions and rotation scheme.

Table 4 indicates the generic skills that are assessed and provides the minimum number of points required to pass, which in total should reach at least a hundred and thirty.

**Table 4. List of clinical skills assessed during PACES across stations with a minimum of points in each station required to pass.**

<i>Clinical Skill</i>	<i>Abdominal</i>	<i>Respiratory</i>	<i>History Taking</i>	<i>Cardiology</i>	<i>Neurology</i>	<i>Communication &amp; Ethics</i>	<i>Clinical Consultation 1</i>	<i>Clinical Consultation 2</i>	<i>Min number of points to pass*</i>
A. Physical Examination	X	X		X	X		X	X	14
B. Identifying Physical Signs	X	X		X	X		X	X	14
C. Clinical Communication Skills			X			X	X	X	10
D. Differential Diagnosis	X	X	X	X	X		X	X	16
E. Clinical Judgement	X	X	X	X	X	X	X	X	18
F. Managing Patients' Concerns			X			X	X	X	10
G. Maintaining Patient Welfare	X	X	X	X	X	X	X	X	28

\* The minimum number of points in individual stations sums to 110, but the overall minimum required by the RCP to pass is 130.

Source: The Royal College of Physicians

If a candidate receives an 'Unsatisfactory' or 'Borderline' mark on any of the skills, it requires a comment from the examiner, as obtaining such marks results in failing the exam. Examiners may provide additional feedback on the mark-sheet as they see fit. However, they are requested to refrain from making any evaluative comments or verbal feedback during the session, and are not allowed to ask any leading questions that might help a candidate. In fact, examiners are purposefully trained and receive detailed guidelines on what is and is not acceptable, in order to maintain high-level of standardisation of PACES. All restrictions aim to prevent even minor violations of the standardisation and fairness of the examination.

### *Reliability and validity of PACES*

The reliability coefficient for PACES was estimated using Rasch modelling by McManus *et al.* (2006) at 0.82. No generalizability studies have been published to date, and generalizability analysis of PACES has experienced a number of technical problems, which mean it is far

from straightforward. A study by Dacre *et al.* (2003) demonstrated that examiners' assessments were highly reliable, which supported PACES being a highly standardised exam. According to that study, only 2.2% of paired markers differed from one another by more than two points (on the old scale) and 97.7% were in agreement within one point. Overall, 60.7% markers were in absolute agreement (Dacre *et al.*, 2003). This was corroborated in the already referenced research by McManus and colleagues (McManus *et al.*, 2006) who showed that 87% of variance in PACES results was explained by differences between the candidates, 1% of difference was variance by station, and 12% of variance was due to examiners bias (the 'hawk-dove effect'). Examiner bias was responsible for an overall 4.1% error in the final Pass/Fail score, with 2.6% of candidates being underscored and 1.5% being overscored. Dewhurst and colleagues (2007) showed that an interaction between ethnicity of candidate and ethnicity of examiners was generally non-significant; however, a small effect was present in the combined communication stations, but not in the clinical stations. Further, the ethnicity of candidates and examiners interaction was non-significant in case of white or mixed pairs of examiners. Non-white pairs of examiners tended to be more lenient towards non-white candidates.

As was the case in Part I and Part II, there were significant differences in candidates' performance depending on their medical school affiliation (McManus, Elder, *et al.*, 2008). According to the study, students from Oxford performed above average, while Dundee, Liverpool, and London candidates below average. Also, significant differences in PACES results were observed for sex and ethnicity (Dewhurst *et al.*, 2007). Women performed better than men, while non-white men performed below expectations in comparison to white men and non-white women. Similar differences were observed in a sample of doctors preparing for PACES in a study by Bessant and colleagues (Bessant, Bessant, Chesser, & Coakley, 2006). Finally, it was also found that candidates who passed PACES were better assessed in all five measures of WBAs, in comparison to those who didn't pass (Levy *et al.*, 2011a).

#### **2.4.2 MRCGP**

The Membership of the Royal College of General Practitioners ('MRCGP') is "an integrated assessment system, success in which confirms that a doctor satisfactorily completed specialty training for general practice, and is competent to enter independent practise in the United Kingdom without further supervision" (RCGP, 2013b). MRCGP is administered by the Royal College of General Practitioners. The MRCGP took its current form in September 2007, aiming to assess twelve key competences required from a general practitioner (Riley,

2008). It now consists of three components: a written Applied Knowledge Test ('AKT'), a Clinical Assessment of Skills ('CSA'), and Workplace Based Assessments ('WBAs') recorded during GP training.

AKT is a three hour test comprising two hundred questions in three question formats: Single Best Answer, extended matching questions, and completion of algorithms (Metcalf, 2012). Approximately 80% of the questions refer to clinical medicine, 10% refer to critical appraisal and evidence based practice, and 10% refer to health informatics and the administrative aspects of practice (RCGP, 2013a). AKT aims to test higher order reasoning and problem solving. Munro and colleagues (Munro *et al.*, 2005) found that the reliability coefficient for AKT varied from 0.85 to 0.88 between 1998 and 2003, and as of 2008, Cronbach's Alpha for AKT was as high as 0.89 (Metcalf, 2012). Further evidence for the quality of AKT is provided by Metcalf (2012).

CSA is a standardised clinical assessment aiming to test a general practitioner's ability in the key domains of data gathering, technical and assessment skills, clinical management skills, and interpersonal skills. Its purpose is to test the ability "to gather information and apply learned understanding of disease processes and person-centred care appropriately in a standardised context, make evidence-based decisions, and communicate effectively with patients and colleagues" (RCGP, 2013b). It is also a goal of CSA to assess ability to successfully integrate the above-mentioned skills. CSA in its current form was introduced in 2007, and replaced the expert assessment of a video documenting candidate's consultation skills and professionalism. The video consisted of five consultations and each was assessed based on presence of fifteen behavioural markers (1 point for each behaviour shown), giving the range of attainable points between 0 and 75 (McKinstry, Walker, Blaney, Heaney, & Begg, 2004). The new format CSA demands candidates examine thirteen cases, which are all simulated general practice consultations (RCGP, 2010) – meaning they are all played by actors. Arguments for this method of testing and evidence for standardisation of actor preparation and case calibration was provided by Russell, Simpson & Rendel (2011). Each consultation lasts ten minutes. Examiners assess candidates on a four-point scale: Clear Pass (3 points), Pass, Fail, and Clear Fail (0 points), and the grades are totalled to a numerical score between 0 and 117 points. Examiners observe candidates and are not allowed to interact with them. Little literature existed on the psychometric qualities of CSA, and recently there was a certain level of controversy on the fairness of CSA towards ethnic minorities (Kaffash, 2012, 2013). For example, it was found that non-white graduates were more likely to fail CSA (Esmail & Roberts, 2013) in comparison to their white UK trained

colleagues. Controlling for AKT results, PLAB tests, and IELTS scores resulted in alleviating the difference in performance between white UK graduates and non-white international graduates, but a potential bias was suggested with reference to UK-graduates from ethnic minorities. On the other hand a study conducted by Wakeford and colleagues (Wakeford, Denney, Ludka-Stempien, & Mcmanus, 2015) suggests that underperformance of non-white candidates in MRCGP does not imply bias, as non-white candidates also perform less well in other independent exams. In fact, in a recent<sup>6</sup> judicial review in the case of the British Association of Physicians of Indian Origin ('BAPIO') against RCGP and GMC a High Court Judge rejected the claims of CSA being racially discriminatory and rejected that RCGP breached its public sector equality duty, by saying, among others, "I am satisfied that the Clinical Skills Assessment is a proportionate means of achieving the legitimate aim identified"(section 45), and identifies that aim in section 43 by stating :

The assessment serves the legitimate purpose of protecting patient safety by means that are, in principle, acceptable to do so at a human cost which is tolerable for those who ultimately succeed. There is no basis for contending that the small number who fail ultimately do so for any reason apart from their own shortcomings as prospective general practitioners.

The Wakeford *et al.* paper (2015) provided estimates for the reliability of CSA; a coefficient of 0.75. However, in the absence of studies providing an estimate of the CSA reliability coefficient at the time when analyses of this thesis were performed, which was prior to the submission of the Wakeford *et al.* paper, the reliability coefficient was assumed to be 0.80.

WBAs are used to evaluate a doctor's progress during GP specialty training in "important psychomotor skills (the clinical and practical skills specific to general practice)" (Riley, 2008, p. 50). Their goal is to assess a GP's ability to gather evidence, provide feedback, and maintain professional behaviour. WBAs are based on standard tools mentioned previously.

According to the current RCGP regulations AKT can be taken only by trainees in the second and third year of training (RCGP, 2012), while CSA is limited to the third-year trainees. As of August 2010 candidates are permitted only four attempts in each part of the examination.

---

<sup>6</sup> Judgment of April 14<sup>th</sup>, 2014 can be found online at <http://www.rcgp.org.uk/news/2014/may/~media/Files/News/Judicial-Review-Judgment-14-April-2014.ashx>

### **2.4.3 Clinical Oncology specialist exam ('FRCR')**

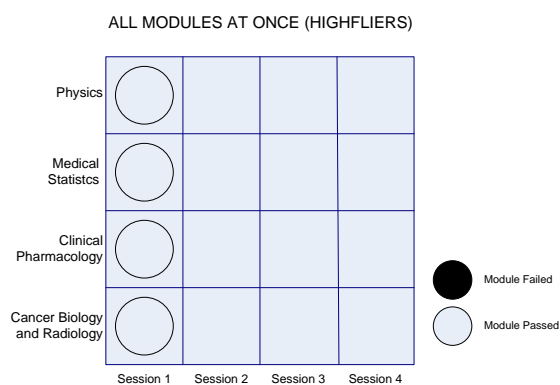
The FRCR exam is a two-stage process and consists of two assessments: the First and the Final exam, called FRCR1 and FRCR2, respectively. FRCR1 is usually taken after admission to Specialty Training in Clinical Oncology and is a written exam consisting of four separate modules: Cancer Biology and Radiobiology, Clinical Pharmacology, Medical Statistics, and Physics (The Royal College of Radiologists, 2011). Cancer Biology & Radiobiology and Physics consist of fifty single best answer questions, while Clinical Pharmacology and Medical Statistics modules consist of forty questions. Candidates are allowed a period of two years to pass all four modules. There are two diets of each module a year, meaning that candidates have a maximum of four attempts at each module. There are no specific rules on the order in which the modules should be taken, but candidates are encouraged to take all modules at once. Candidate choices do, however, vary and the strategies for passing FRCR1 are presented in Figure 7.

Candidates can attempt all modules at once and pass on their first attempt (Example 1). Others might take modules one at a time (Example 2) and pass them on first attempt, or may need more attempts in any of the modules (Examples 3 and 4). Candidates can drop out after attempting only some of the modules (Example 5 in Figure 7 above), while some other candidates may ultimately fail in one or more of the modules within the two-year timeframe, which excludes them from the specialty (Example 6).

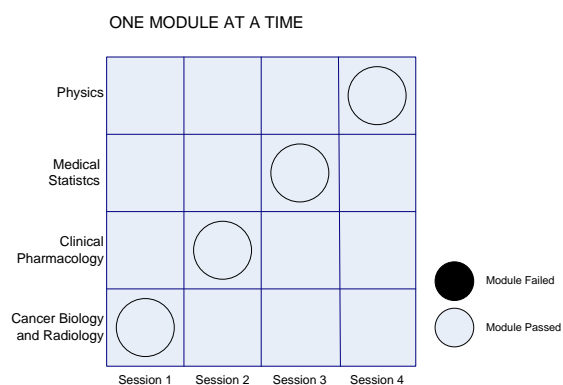
FRCR2 is usually taken after completing three years of supervised Clinical Oncology training, and after passing all modules of FRCR1 (RCR, 2013c). The emphasis of this part of the exam is on radiotherapy and drug therapy; however, a good general medical knowledge is also expected of candidates. FRCR2 is taken in two parts: Part A and Part B. Part B is required to be passed within five sittings from the date of passing Part A (RCR, 2013d). Part A consists of a written exam in two papers, each being a hundred and twenty Single Best Answer questions long (RCR, 2013a). The Single Best Answer format was introduced in 2006, replacing True/False MCQs (Tan & McAleer, 2008). Part B comprises a clinical and oral examination. The clinical examination is divided into five stations, each eight minutes long. Performance at each station is assessed by two examiners (RCR, 2013b). The oral examination lasts forty minutes in total and is divided into two stations with two pairs of examiners. Each examiner asks questions for ten minutes, while all examiners assess the candidate's performance. A summary mark is calculated based on an overall performance in both parts of FRCR2. Until Autumn 2010 the components of FRCR2 were marked in a form of bands from A to F, while in 2011 the marking was changed to numerical scores: in

the Spring session it was a raw numerical score; in Autumn the score was given as a percentage.

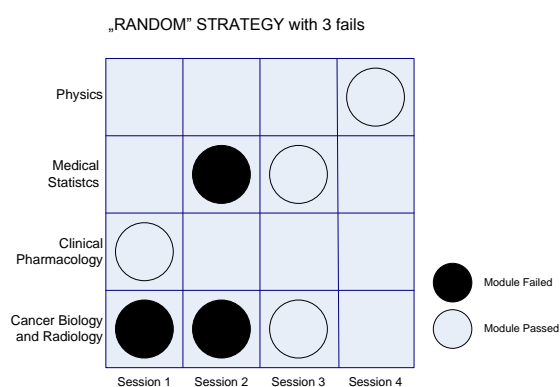
### Example 1 – Pass



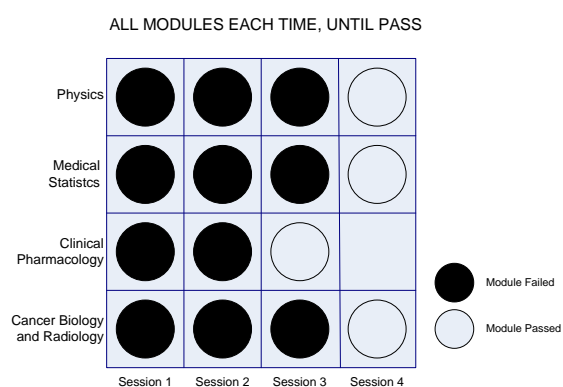
### Example 2 – Pass



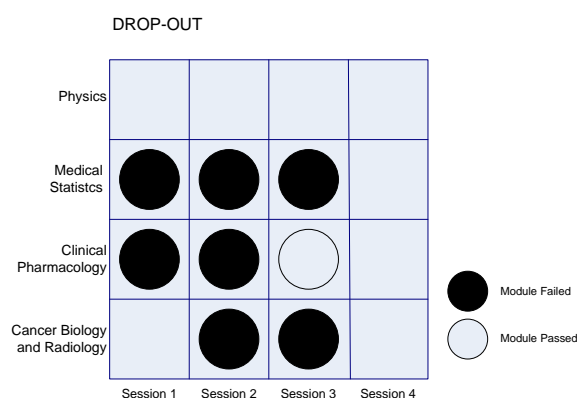
### Example 3 – Pass



### Example 4 – Pass



### Example 5 – Fail through dropping out



### Example 6 – Fail

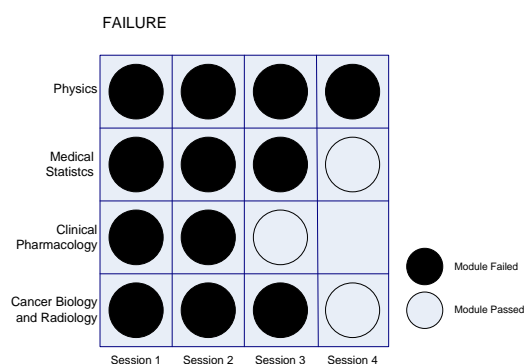


Figure 7. Possible outcomes and strategies in attempting FRCR1.

Little published information exists on the psychometric quality of the FRCR exams, and the existing literature focuses on FRCR2. As indicated by Tan and McAleer (2008), the change in FRCR2 written test format increased the reliability of the examination. However, the paper does not provide exact coefficients. Further, as assessed by Yeung and colleagues, the FRCR2 B oral examination was moderately reliable (Yeung, Booth, Larkin, McCoubrie, & McKnight, 2012), whilst candidates evaluated it as fair (Yeung, Booth, Jacob, McCoubrie, & McKnight, 2011). For the purposes of this research the RCR provided the unpublished reliability coefficients for the written parts of the FRCR1 and FRCR2 as Kuder-Richardson K20 coefficients (Anastasi & Urbina, 1997). In 2011 the reliability coefficient for FRCR1 modules reached on average 0.73, while the reliability coefficient for FRCR2 written paper was 0.85.

#### **2.4.4 Specialty Certificate Exams ('SCEs')**

The SCEs are examinations developed by the RCPs in cooperation with the Specialist Societies. They are administered by the RCPs and aim to test specialist knowledge of trainees in twelve different specialties: Acute Medicine, Dermatology, Endocrinology, Gastroenterology, Geriatric Medicine, Infectious Diseases, Neurology, Medical Oncology, Palliative Medicine, Renal Medicine, Respiratory Medicine, and Rheumatology. SCEs are compulsory assessments required for obtaining a CCT for all UK trainees whose specialist training began in or after August 2007 (RCP, 2013e). SCEs are usually attempted in the penultimate year of training, approximately four to six years after the completion of MRCP(UK). An expected time span for completing one specialty is eight to nine years from the day of PMQ. Some physicians are certified in more than one specialty, which means they have passed two or more SCEs; however, this is the case for only 0.7% of all registered physicians, based on the List of Registered Medical Practitioners (GMC, 2011).

An SCE is a knowledge assessment in the form of a computer-based test<sup>7</sup>. It comprises two three-hour papers of a hundred questions in the Best of Five format each; two hundred questions in total. The questions aim to assess the ability to interpret clinical information and solve problems (RCP, 2013f). Each SCE can only be attempted once a year. As of July 2011 Dermatology and Geriatric Medicine require that candidates hold MRCP(UK) diploma.

---

<sup>7</sup> The facilities for examinations are provided by PearsonVue, which is a commercial company across UK. More can be found on <http://www.pearsonvue.com/>.



Little literature exists on the quality of the SCEs. Cookson (2010) suggested that even though SCEs were acknowledged to be sufficiently reliable (although inconsistently) based on the pilot studies of 2006, arguments for the validity of SCEs were lacking. He also pointed out that the question base for pilot studies was limited and that SCEs assessed only knowledge. Those concerns were, however, addressed by Dacre and Mucklow (2010), who argued that the SCEs question base had been expanded since pilots, and that the SCEs were designed purposefully to test for knowledge and not clinical skills. They have also raised the issue of reliability and argued that the case for SCEs was similar to the one of Part II (Tighe *et al.*, 2010): in a limited sample of doctors with similar ability (i.e. with low variance) in a test of two hundred questions the reliability is unlikely to reach 0.90. At the same time, they pointed out that for nine out of eleven SCEs the reliability coefficient reached 0.8. Based on the unpublished materials provided by the RCPs for the purposes of this research, since Cookson's publication the reliability coefficients increased, and in 2011 ranged from 0.79 to 0.94 depending on the specialty. Mean reliability coefficients from all years the examinations existed for, for each specialty reached: 0.83 for Acute Medicine, 0.87 for Dermatology, 0.89 for Endocrinology, 0.82 for Gastroenterology, 0.76 for Geriatric Medicine, 0.94 for Infectious Diseases, 0.90 for Neurology, 0.83 for Medical Oncology, 0.85 for Renal Medicine, 0.84 for Respiratory Medicine, 0.90 for Rheumatology, and 0.82 for Palliative Medicine.

#### **2.4.5 Cardiology specialist exam: Knowledge Based Assessment ('CKBA')**

The Cardiology Knowledge Based Assessment ('CKBA') is a computer-based exam developed jointly by the British Cardiovascular Society and the European Cardiac Society. It aims to assess knowledge in the core areas of cardiovascular medicine for cardiology specialty trainees. Candidates are expected to attempt it in the penultimate year of training (ST5), which means it takes place 4 to 6 years after MRCP(UK). As of 2007, CKBA is a mandatory part of the final specialty assessment resulting in awarding a CCT in cardiovascular medicine. Alongside WBAs, CKBA is considered a part of the Annual Review of Competence Progression assessment in the final year (BCS, 2013b). This means that passing or failing CKBA may effectively delay the completion of training. CKBA design was modelled on Part II of MRCP(UK); therefore, it employs applied clinical knowledge questions. CKBA consists of a single three-hour paper with a hundred and twenty multiple-choice questions (BCS, 2013a).

A study on the CKBA pilot in 2009 indicated that the exam is sufficiently reliable for a high-stakes examination (BCS, 2010) with Cronbach's Alpha reaching 0.75.

#### **2.4.6 Annual Review of Competence Progression ('ARCP')**

The ARCP is a process aiming to provide feedback to trainee doctors on their training progression. The feedback is based on three elements: educational appraisal, assessment, and planning, as described in the *Gold Guide* (NHS, 2010a). ARCP is a formative process, and is based on the information gathered by the trainees in their e-portfolio. For the ARCP panel to make a decision on a trainee's progress the following evidence must be provided: Workplace Based Assessment forms, exams information, additional training information, a reflective logbook (e-portfolio), a structured report from the educational supervisor, and a Personal Development Plan for the year. The ARCP Panel makes a judgement on the progress against the required competences for CMT or ST. The direct result of the assessment procedure is that while a trainee is in CMT, an ARCP outcome is dependent on them passing or failing MRCP(UK). The review takes place at least every twelve months; however, upon certain outcomes additional reviews might be scheduled. There are nine possible outcomes of the panel review (NHS, 2010a):

**Outcome 1: Satisfactory Progress.** Development of competences progressing at the expected rate.

**Outcome 2: Development of specific competences required with no additional time for training required.** This outcome means that the progress was acceptable, however, certain competences were not fully achieved in the year of assessment. The trainee needs to provide evidence that the requirements were met at the next review.

**Outcome 3: Inadequate progress - additional training time required.** Obtaining this review outcome means that additional training time is required; the training time is formally extended, usually by a year.

**Outcome 4: Released from training programme, with or without specified competences.** The recommendation of release is given when there is an insufficient progress despite assigning additional training time.

**Outcome 5: Incomplete evidence presented – additional training time may be required.** This outcome results from insufficient evidence, missing or no information on the progress. The trainee may deliver missing evidence with explanation.

**Outcome 6: Gained all required competences.** This outcome means that the trainee has completed the training and is recommended for an award of CCT.

**Outcome 7: Fixed-Term Specialty Trainee.** This outcome is assigned to trainees who either decided to participate in fixed-term training or are undertaking additional training within a training programme. It divides into four possible outcomes: satisfactory progress, development of specific competences needed, inadequate progress, and incomplete evidence presented.

**Outcome 8: Out of programme for research, approved clinical training or a career break.** This outcome is reserved to those candidates who decided to pursue other paths of personal development, such as research, or who decided to have a break from their medical career. It is possible that if such a 'break' in the formal training is contributing to gaining competences required by the trainee's programme, then the documents from this period can be taken into account by the panel in the usual way.

**Outcome 9: Undertaking top-up training in a training post.** This outcome is reserved for those doctors who applied for a position in an approved specialty training for a limited period of time. However, those individuals should then submit in-work assessments and documentation so that the panel can later assess if the GMC recommended objectives have been met by the trainee.

Effectively, outcomes 1 to 6 are typical and most prevalent, while outcomes 7 to 9 should be reviewed qualitatively, as they signify an individual path of career development.

#### **2.4.7 List of Registered Medical Practitioners ('LRMP')**

The LRMP is a list of all doctors registered to practise medicine in the UK. It is administered by the GMC, and was first established in 1858. The purpose of the LRMP is to provide accurate up-to-date publicly accessible information on the status of medical practitioners. Maintaining the LRMP is a part of the quality assurance process in the UK healthcare system. Registered doctors are required to abide by the standards set by the GMC. As of 2009, in order to be able to treat patients, doctors need to be not only registered with the GMC, but are also required to hold a licence to practise (GMC, 2009). Doctors can be additionally listed on either the GP Register or the Specialty Register, depending on their area of practice; however, this is only possible after they have completed their training (GMC, 2013b). Apart from full registration which signifies a licence to practise, provisional registration (for Foundation Year 1 doctors), and registration without licence (for researchers and non-practising doctors), the LRMP reflects all sanctions imposed on a doctor by the GMC. Such sanctions include limitations to the right to practice: erasures, suspensions, warnings, conditions, and undertakings imposed on a doctor due to

impairment of their Fitness to Practice (GMC, 2012c), as described earlier. The list of the twelve categories of registration and their definitions are presented in Table 5.

**Table 5. List of registration statuses on the LRMP with explanations.**

<i>Status</i>	<i>Meaning</i>
(1) Registered with Licence	Registered with full licence to practise.
(2) Provisionally Registered with Licence	Usually reserved for Foundation Year 1 doctors. Allows practise only in approved posts.
(3) Registered without Licence	Registration for doctors who do not wish to practise medicine.
(4) Provisionally Registered without Licence	Status for doctors who were provisionally registered, but relinquished their licence.
(5) Administrative Erasure	Removed for various administrative reasons not related to practise; often resulting from a failure to pay fees or due to lack of response to GMC correspondence.
(6) Relinquished	Voluntary erasure from the register, which may be due to performance issues or other personal reasons, e.g. decision to practise medicine abroad. Registration can be restored if the reasons were not related to fitness to practise.
(7) Deceased	Removal after death.
(8) Erased after Fitness to Practise review	Removed from the Register after being judged unfit to practise. Erasure lasts from 5 years to lifetime.
(9) Suspended	Temporary status that requires the doctor to refrain from practising medicine within a certain period of time, as a result of a Fitness to Practise review.
(10) Received a Warning	Concern about a significant departure from the principles of Good Medical Practice, but fitness to practise not impaired
(11) Undertakings	Doctor holds a licence to practise but needs to undertake specific steps to address the issues that raised concern, usually agreed with the doctor. Can be removed if there is evidence that they are no longer needed.
(12) Conditions	Doctor holds a licence to practise under certain conditions imposed by the Fitness to Practise Panel, e.g. supervision or restriction to certain fields of medicine. Conditions last up to three years.

Source: based on the GMC website

Until 2004 the LRMP was published on paper ('The Medical Register'), and for some years before that it was also available as a CD-ROM. From 2005 it has only been available in an electronic version on the GMC website. The LRMP is updated daily, and it can be

downloaded every day by those who subscribe to it. Historical versions of the LRMP are not available, files only being available for seven days after their release. On September 4<sup>th</sup>, 2014 the LRMP held information on 267,503 doctors, of which 212,922 were under full registration, and 8,116 were under provisional registration (GMC, 2014a).

The heed which the GMC pays to updating the LRMP means that the register can be assumed to be an accurate reflection of the current status of all medical professionals in the UK, and therefore, that the data are robust. However, it needs to be pointed out that LRMP status categories are quite general; there is no distinction between individual doctors within each category, which results in a low variability of any measure based upon such categories.

#### **2.4.8 Investigation by the GMC (Fitness to Practice review documentation)**

As referred to above, Fitness to Practice procedures are the GMC and MPTS investigations after a complaint into whether a doctor is fit to practise medicine. There are three stages of the Fitness to Practice review process ('GMC FtP'). At the initial stage, enquiries are categorised into three groups: those that cause a serious concern ('Stream 1'), those that may not be serious but potentially could pose a risk to patients if a wider pattern was revealed ('Stream 2'), and those that are not pursued after an initial investigation (GMC, 2012a). Stream 1 was therefore of main interest in relation to the purpose of this thesis. In the first phase of Stream 1 an enquiry may be closed or followed through to the second phase when it is investigated by a GMC case examiner. The examiner may then either impose limitations to the doctor's licence, close the investigation, or progress the inquiry to the final stage, which is a formal assessment of competence. These formal examinations are adjusted to the area of expertise of the referred doctor and they do not have a set pass-mark but rather they are set against a benchmark.

Although tailored to fit a specific case these assessments of performance aim to assess knowledge, clinical skills, and professional behaviour. For example, a GP would need to take a knowledge test, a simulated surgery task, and an OSCE. The knowledge tests last two hours and consists of a hundred and twenty Single Best Answer questions that refer to the GP practice. The simulated surgery task comprises ten surgery consultations that are typical for the GP setting and likely to occur on a normal working day. Each simulation lasts approximately ten minutes. In the OSCE assessment a doctor is presented with twelve scenarios typical for their work, each in a form of a station (similar to PACES, above). These

cases may involve patients, actors, and interaction with members of medical staff. The surgery and OSCE tasks aim to assess clinical and communication skills and professionalism.

The assessment is made based on a benchmark score. For the knowledge assessment the performance is based on the scores obtained by a reference group – being volunteer doctors of the same specialty – and through comparison with a standard set mark established using Angoff's method (Maurer, Alexander, Callahan, Bailey, & Dambrot, 1991). The simulated surgery assessment is more complicated. In general it is made through a comparison with the scores of the reference group, but the communications skills are assessed by a lay person. The clinical skills are assessed in view of the *Good Medical Practice* (GMC, 2013a) by all of the clinical assessors. The result is compared with the standard set mark using modified contrasting groups method (Burrows, Bingham, & Brailovsky, 1999).

After the process is complete the entire case is examined by the Fitness to Practise panel who decide if action is required on a doctor's registration. The choice of sanctions is limited and consists of the above-explained conditions, undertakings, warning, suspensions, and erasures. Apart from an erasure, which results in an irrevocable removal from the register, the other licence actions are time-limited. For example, a doctor can have conditions imposed on their licence for a certain time, e.g. three years, after which their performance and registration restrictions are reviewed and may be lifted.

As mentioned, the GMC maintains their own register of cases of doctors under review; however, the data secured for the purposes of this research may not have included a complete record of investigated cases.

## **SUMMARY**

This chapter presented the sources of criteria for the predictive validity study of MRCP(UK). The choice of those sources was dictated by four key factors. The first factor affecting the choice were the assumptions underlying the construction of the MRCP(UK), which state that MRCP(UK) measures the skills, knowledge and attitudes. The second factor was that the criteria used for the purposes of this research would be quantitative rather than descriptive. The third aspect while making a choice was that the criteria should be embedded within the UK medical education system. The fourth factor was the quality and availability of data sources. It was also taken into account that in order to avoid common

method variance bias in making validity inferences these criteria should originate from as many sources as possible, and should be as diverse as possible.

Several suitable sources of validity data were found. However, based on the availability and quality of the data sources the selection was limited. The knowledge criteria were based on the MRGP AKT exam, FRCR exams, Specialty Certificate Exams and Cardiology exam. It was decided that clinical skills and professional attitudes would be measured jointly using MRGCP CSA results, the FRCR2 clinical component, the Annual Review of Competence Progression assessment outcomes, the GMC LRMP registration data, and the Fitness to Practice records. Being erased from the LRMP for administrative reasons was used as a proxy for conscientiousness, which was the sole separate criterion for assessing attitudes.

The above sources for the various criterion measures have been described, and evidence for their reliability and validity was provided where possible to support the robustness of the subsequent analyses. Information on the factual contents of the datasets that were further obtained, and the description of measures used in analyses are provided in Chapter 3. The following chapter also provides detailed information on the process of merging the data and other methodological aspects of this study.

## Chapter 3

### Methodology

#### ABSTRACT

*Following the presented description of the UK medical education setting, this chapter presents the methodology of investigating the predictive validity of MRCP(UK). The research assumed a retrospective longitudinal approach. Random sampling was not employed and the effective sample sizes for analyses varied from 5 to 25,447 cases, depending on the criterial dataset and MRCP(UK) part. The main file was a dataset with 50,311 records of MRCP(UK) candidates who attempted MRCP(UK) between May 2003 and January 2011. Other materials used in this study are datasets containing results of exams described in Chapter 2, the list of registered medical practitioners in the UK, records of ARCP training outcomes for physician trainees, and the list of GMC fitness to practice procedures, provided by institutions responsible for their administration. The procedure of merging the datasets with the main MRCP(UK) file was based on the GMC number or RCP Candidate number. A description of the actual measures with the full list of variables employed in the analyses is provided. The issues of missing data are addressed, followed by an overview of the statistical procedures, and statistical software used for the purposes of this research. The limitations of the study are discussed, and the ethical approval confirmation is provided.*

#### 3.1 SAMPLE AND SAMPLING TECHNIQUE

The sampling method was judgment sampling, which is a non-random sampling technique. This means that the participants for the study were not chosen at random, but rather specifically for the purposes of the particular project. In the case of this research, participants were candidate physicians, who attempted MRCP(UK) between May 2003 and January 2011, but who also attempted any of the exams described in Chapter 2, had any ARCP records, were registered with the GMC, or were investigated by the GMC FtP panel. Therefore, the selection procedure was based on judgment and not on drawing from the population. The main MRCP(UK) dataset ('History File') which is thoroughly described later, initially contained 50,311 cases of doctors attempting MRCP(UK) within the defined time-frame. However, due to the merging procedures, the analyses were performed on several smaller separate samples. The size of each sample was dependent on three factors: the size of the criterial dataset provided for this research, the number of the records that could



have been matched due to timeframe overlap of the datasets, and the existence of the key identifier (the GMC registration number or the RCPs candidate number). Further, the sample sizes were also limited due to other specific factors. In the particular case of MRCGP the number of analysed records was further limited due to the chronology of the exams. The nature of this study indicated that the records suitable for analyses were only those where MRCGP was taken after MRCP(UK); therefore, all cases where either AKT or CSA were taken before MRCP(UK) were excluded. In the case of ARCP data a significant number of valid cases had to be excluded due to the fact that some assessments were provided during the Core Medical Training and such outcomes are dependent upon passing MRCP(UK). If it had been included in the analyses, the criterion based on the ARCP outcome would have become contaminated.

Table 6 summarizes the nominal dataset sizes and effective sample sizes for each of the criteria and for each of the MRCP(UK) parts.

**Table 6. Nominal and effective sample sizes for all datasets used in current research.**

<i>Criterion</i>	<i>Dataset cases</i>	<i>Matched records</i>	<i>Part I effective sample size</i>	<i>Part II effective sample size</i>	<i>PACES effective sample size</i>
<b>Knowledge assessments</b>					
MRCGP AKT	7,685	1,976	1,976	938	739
FRCR1	1,032	746	228	292	402
FRCR2 written exam	352	339	80	147	246
SCEs	2,244	2,076	1,466	1,859	2,063
Cardiology	209	209	123	184	209
<b>Clinical skills assessments</b>					
MRCGP CSA	7,685	1,976	1,976	938	739
FRCR2 clinical and oral	352	339	80	147	246
<b>Overall performance</b>					
LRMP	326,822	33,359	25,447	18,757	18,760
ARCP	6,306	2,979	2,539	2,836	2,916
Under GMC investigation	820	8	5	6	8

## 3.2 MATERIALS

Materials for the research comprised the datasets obtained from RCPs, RCR, GMC, BCS, and JRCPTB. All data used in the research were pre-existing data and no additional information was collected during the course of this study. In all cases the administering bodies were responsible for the data collection process and data accuracy.

### 3.2.1 MRCP (UK) Dataset details

The dataset provided by RCPs contained 50,311 individual records of candidates attempting MRCP(UK) between May 2003 and January 2011. It held information on individual percentage results from different parts of MRCP(UK) which were later converted into standardised scores with the pass-mark set as zero for ease of reference during analyses. Results were provided for all attempts in each of the three parts. The total number of attempts for all candidates, the pass-marks for each examination sitting (also called 'diets'), and the demographic data: sex, self-declared ethnicity, date of qualification, cohort of qualification, date of birth, current age, PMQ, and being a probable UK trainee were also provided.

The time-frame for this dataset was chosen on purpose; during this period MRCP(UK) did not change extensively, which allowed for the analyses to be based on stable data. Minor changes to the exam included:

- Part I was modified in May 2003 so that the questions were written in the Best-of-Five format, which is why the start date of the dataset was chosen on this diet.
- Part II changes included varying number of marked questions from diet to diet. As referred to above in Chapter 2, the number of questions was increased in order to obtain the appropriate level of reliability; an issue discussed in Tighe *et al.* (2010). As the results of Part II were provided as a percentage, the actual number of questions that were marked (which varied slightly from diet to diet) was not factored into the analyses, nor was any change in reliability of Part II results;
- In 2009 the components of PACES changed; in the provided dataset all results prior to 2009 were re-scaled to the new marking system. The provided dataset contained consistent marks. Therefore, all analyses were also performed on consistent data.

### 3.2.2 MRCGP Dataset details

The dataset provided by RCGP contained information on 7,685 individual doctors attempting MRCGP examination between October 2007 and May 2011. The dataset

contained information on sex, year of qualification, self-declared ethnicity, and first attempt scores for AKT and CSA. The scores were provided as standardised against the pass-mark (with zero being the pass-mark), binomial (in Pass/Fail format), and in the case of CSA also in equated scores standardised against the pass-mark. The equated scores were made available for this study due to the changes made to CSA in 2007. The statistical equation procedure was performed by the RCGP. The predictive validity analyses were performed separately for the old marking scheme, new marking scheme, and the equated scores. No information on WBAs results was included in the data. As mentioned previously, due to the nature of this predictive validity study only those candidates who attempted MRCGP after MRCP(UK) were taken into account during analyses.

### **3.2.3 FRCR Dataset details**

The RCR provided two datasets containing information on the FRCR (Clinical Oncology) specialty exams separately: FRCR1 and FRCR2. Additionally, RCR also provided paper records with detailed results of FRCR2. The FRCR1 dataset contained information on 1,032 candidates attempting exams between March 1997 and January 2011. These dates were chosen as broadly as possible in order to maximise the number of records linked to the MRCP(UK) database. The limitations in the numbers resulted only from the availability of the electronic records for FRCR1. The FRCR1 dataset contained information on the number of attempts in each of the FRCR1 modules with information on whether a candidate passed or failed the module. For records post-2006 also additional information on the actual raw mark in each module was included.

The FRCR2 dataset initially contained information on a total of 352 candidates who attempted the exam between 1990 and January 2011; however, much of the data prior to 2005 were incomplete due to unknown<sup>8</sup> technical reasons. This effectively limited the data. The provided dataset contained 339 linkable records. Information included in that dataset was analogous to the FRCR1 dataset: the demographic data, the number of attempts, and the final pass/fail score. The information in that dataset was further amended after obtaining paper records for partial scores in FRCR2 written, oral, and clinical modules. Between 2004 and 2006 there were two written papers in the FRCR2, while later there was just one. Therefore, the grades from the two written papers were averaged for each candidate into one grade to unify the marking system. Initially, for diets Spring 2004 to

---

<sup>8</sup>At the completion of this PhD thesis those reasons were still investigated by the administering body; the most plausible explanation to date was database error.

Autumn 2010 the scoring system assigned letter grades from A to F (fail). Scores were provided as total points gained for the Spring 2011 diet, but as percentages for the Autumn 2011 diet. Scores from all three parts of FRCR2 were merged into the dataset as Z-transformed scores (with a mean of zero and a standard deviation of one) due to differences in the marking schemes across diets. The scores were standardised within each diet separately.

### 3.2.4 SCEs Datasets details

Twenty six SCE files contained 2,563 individual scores of 2,244 doctors attempting at least one of the SCEs between 2008 and 2011. The accurate numbers of doctors for each specialty are presented in Table 7.

**Table 7. The overall number of records from the SCE files with the list of years of examinations taken into account.**

Specialty	Years of Exams	Number of records
Acute Medicine	2010, 2011	251
Dermatology	2009, 2010, 2011	178
Endocrinology	2009, 2010, 2011	352
Gastroenterology	2009, 2010, 2011	413
Geriatric Medicine	2009, 2010, 2011	330
Infectious Diseases	2009, 2010, 2011	55
Neurology	2009, 2010, 2011	167
Medical Oncology	2010, 2011	124
Renal Medicine	2009, 2010, 2011	211
Respiratory Medicine	2009, 2010, 2011	290
Rheumatology	2010, 2011	154
Palliative Medicine	2011	38
<b>Total:</b>	--	2,563

Source: The Royal College of Physicians

Each file contained raw scores for each exam question, total raw score, percentage score, pass-mark for each year, and pass/fail information for each specialty for a particular year. It also contained demographic data: university of PMQ, year of PMQ, date of birth, sex, self-declared ethnicity, if a candidate was a UK trainee or not, and the RCP candidate number. The files also provided information on the location of the examination centre where the exam was taken. The twenty-six files were merged into one long file and restructured so

that each doctor constituted one case in the dataset. The final restructured dataset contained 2,224 records.

### **3.2.5 Cardiology Dataset details**

The dataset contained information on the first attempt result in the CKBA exam for 209 doctors attempting it in 2010 and 2011. The results were recorded as a percentage score. The file also included the raw number of responses with correct answers, year of examination, and the pass-mark in that year. It also contained the RCP candidate number.

### **3.2.6 ARCP Dataset details**

The dataset was provided by JRCPTB and contained 10,610 records on 6,306 individual specialty trainees under supervision of JRCPTB, including Cardiology trainees. The dataset contained the names and the GMC numbers of trainee doctors, along with information on responsible deanery, stage of training, year of training, and place of training, together with up-to-date ARCP outcomes with a qualitative justification. The file contained multiple entries for some candidates, which provided longitudinal information on their assessments' outcomes. For the purposes of the analysis all records of the CMT trainees were excluded, as explained previously, in section 2.4.6. Only the final and most complete record for each trainee was analysed.

### **3.2.7 LRMP Dataset details**

The LRMP dataset was an aggregate of fifty files downloaded from the GMC website between September 2008 to November 2012. Each file contained a full list of up-to-date registered doctors with their name, GMC number, and current status of registration. A specially designed Matlab programme written by IC McManus scanned those files to search for any changes in the LRMP status. The final dataset contained 326,822 records of all doctors listed in the LRMP, their GMC number and a notification ('flag') of any registration status that occurred for each doctor within the time-frame of four years. Twelve statuses were flagged, as listed in the Table 8 (next page).

Table 8 also contains frequencies of each status as counted by the programme. The most prevalent were flags for being registered with the licence to practise (97.6% of the sample) and voluntary erasures (11.7%), followed by administrative erasures (9.7%). For technical reasons (due to memory limitations), the version of the programme used here did not count occurrences of each status for each doctor, the order in which they appeared, or the length of the period during which each status was valid.

**Table 8. Frequency of LRMP registration status occurrences over the 4-year period of time for which the data was collected.**

<i>Status</i>	<i>Licence Issues</i>	<i>Count</i>	<i>%</i>
(1) Registered with Licence	No	319,139	97.6%
(2) Provisionally Registered with Licence	No	15,253	4.7%
(3) Registered without Licence	No	19,835	6.1%
(4) Provisionally Registered without Licence	No	109	~0.0%
(5) Administrative Erasure	No	31,117	9.7%
(6) Relinquished	No	38,230	11.7%
(7) Deceased	No	6,059	1.9%
(8) Erased after Fitness to Practise review	Yes	516	0.2%
(9) Suspended	Yes	1,456	0.4%
(10) Received a Warning	Yes	1,169	0.4%
(11) Undertakings	Yes	701	0.2%
(12) Conditions	Yes	1,451	0.4%

Source: LRMP

### 3.2.8 Fitness to Practice Dataset obtained from the GMC

The Fitness to Practice dataset contained 820 records of doctors who were under GMC investigation between 1998 and 2008. The dataset contained names of doctors, their GMC numbers, reason(s) for complaint, GMC and FtP panel comments, results of assessments, and the final outcome. Due to the high sensitivity of the data, the only information extracted from the file were the GMC numbers of those who were investigated. Effectively the matching of such a list and the History File yielded a binary variable which showed who was under review *versus* who was not under review by the GMC.

## 3.3 CRITERION MEASURES AND VARIABLES

In the previous chapters words such as criterion measures, criteria, and source of criterion were used almost interchangeably. However, their consideration should be more focused and so should be their use. Therefore, in order to clarify the distinction, the examinations (e.g. MRCGP) and processes (e.g. licence registration) described in Chapter 2 are regarded as the sources of criterion measures. Performance in an exam, for example, is a criterion that represents a construct of either knowledge or clinical ability/professional attitude. For example, *performance* in the FRCR2 clinical examination is one of the measures of *clinical*

*ability*. Hence, ‘performance’ is the criterial measure, while ‘clinical ability’ is the construct. The variables quantify the performance, and therefore, quantify the criteria. For example, *first attempt scores* in FRCR2 clinical assessment are a numerical expression of the *performance* in that exam. Since the sources of criterion measures were already described in Chapter 2, this section provides details on the measures and the variables used in this study.

### **3.3.1 Measures**

#### **3.3.1.1 MRCP Measures**

Due to the design of MRCP(UK) it was assumed that the first attempts scores in all three parts of MRCP(UK) would be the measures of performance. First attempt results were considered unbiased in terms of learning the form of the exam, and as indicated in the paper by McManus and Ludka (2012) the first attempt scores were predictive of all other attempt scores. Other measures of performance such as the number of attempts and Pass/Fail outcomes were analysed for descriptive purposes, but not used for making inferences. Another measure of performance was passing all MRCP(UK) parts at first attempt. The group of such candidates was named the MRCP(UK) Highfliers, and the fact of being a MRCP(UK) Highflier was used as a factor in contrasting groups analyses.

Part I and Part II were assumed to represent the knowledge component, while PACES was assumed to measure clinical ability and appropriate attitudes and behaviours. Part I, Part II and PACES first attempt scores were used in a standardised form, either against the pass-mark, or as Z-transformed scores, depending on the type of analyses. For example, for the purposes of regression models Z-scores were used. Detailed information on MRCP(UK) was provided in section 2.4.1, and the analyses of the relationships between MRCP(UK) parts, preceding the validity analyses, constitute Chapter 4.

#### **3.3.1.2 MRCGP Measures**

MRCGP consists of two exams: AKT and CSA. The first attempt results in AKT standardised against the pass-mark were used as measures of knowledge, while first attempt results in CSA standardised against the pass-mark were assumed to be a measure of clinical performance. The dataset also provided partial scores in AKT subscales: Organizational Questions, Clinical Medicine, and Evidence Interpretation. Only first attempt scores were employed. The CSA scores were provided in three schemes: the new format, the old format, and the equated format, and all three scores were used in the analyses. The scores

were either standardised against the pass-mark, or for the purposes of comparison between the regression models, Z-transformed.

### **3.3.1.3 FRCR Measures**

Due to the complex structure of the FRCR examination a variety of measures were employed: the mean score on the first attempt in FRCR1 modules, the first attempt scores in separate modules of FRCR1, the standardised first attempt results in the written, oral, and clinical components of FRCR2, and the FRCR2 pass/fail score. Additionally, the variety of strategies for passing FRCR1 (as described in section 2.4.3) and their final outcome allowed for categorising candidates into:

- Highfliers: a group of candidates who passed all modules in one session (Figure 7, Example 1)
- Typical: a group of candidates who either needed more sessions than just one, or had more than one attempt in any of the modules (Figure 7, Examples 2-4)
- Dropouts: a group of candidates who did attempt the exam but either did not attempt one of the modules at all within their two-year window, or did not take any modules in their last session (Figure 7, Example 5)
- Failed: a group of candidates who used all four diets and attempted all four modules, and failed (Example 6).

This classification into groups is hereafter referred to as the 'FRCR Rank'.

As a result of missing FRCR1 scores before 2006, an additional proxy measure of performance in FRCR1 was devised: the number of attempts in FRCR1 modules. The Total Number of Attempts variable was based on the procedure of passing FRCR1, and it ranged from one to sixteen (four sessions x four modules). As described in Chapter 2, each candidate is permitted up to four attempts in each of the four modules. The best candidates needed just one attempt in each module (i.e. four attempts in total), while less well performing candidates required more attempts to pass (five to sixteen). Therefore, all candidates had a score between one to sixteen attempts, with four being the best result, and whenever the total number of attempts was one to three it meant that a person dropped-out or was 'censored' (censoring is described more thoroughly in section 3.5.1).

All measures devised based on FRCR1 together with the first attempt scores in the written and oral part of FRCR2 were considered representative of medical knowledge. The clinical



component of FRCR2 was related to clinical performance. The binomial FRCR2 Pass/Fail score was perceived as a general measure of performance.

#### **3.3.1.4 SCEs and CKBA Measures**

The SCEs and CKBA are written knowledge tests and were assumed to represent the knowledge component of professionalism. The first attempt scores were standardised within each specialty against the mean, and were employed as a measure of assessment of future knowledge. The standardisation of the results within each specialty (including CKBA) was necessary to equate the results of exams. There were no other means of comparison between them: there were no common anchor questions, the specialty exams differed in terms of distribution of the results, and each had a different pass-mark from year to year. The results were also Z-transformed for the purposes of comparison between the linear regression models.

#### **3.3.1.5 ARCP Measures**

The standard ARCP outcomes were divided into satisfactory progress (outcomes 1 and 6) and unsatisfactory progress (outcomes 2 to 5), which dichotomised the criterial measure. Outcomes 7, 8 and 9 were omitted as they describe a situation that does not apply to a standard training programme. For each candidate in the dataset a total number of positive and negative outcomes throughout training was calculated, and then a binary variable ('Overall Progress Assessment') was coded: '0' if all assessments obtained within the programme were positive (which will be referred to as Satisfactory Progress), and '1' if any assessment resulted in an outcome 2, 3, 4 or 5 (Unsatisfactory Progress). This variable was considered a measure of physicians' on-the-job performance or clinical ability and attitude.

Due to the fact that ARCP assessment for CMT trainees is dependent on passing MRCP(UK), all ARCP outcomes taken during CMT were excluded from the analyses.

#### **3.3.1.6 LRMP Measures**

The registration statuses described in section 3.2.7 were coded for all doctors in the LRMP dataset in a form of binary flags. Warnings, conditions, undertakings, suspensions and limitations of licence were recoded to create an aggregate measure of the Licence Issues (also in a binary form). The decision to dichotomise the LRMP measures stemmed directly from the fact that LRMP does not provide information on the reasons of imposing a particular Licence Issue. Further, the computational limitations of the programme mentioned in section 3.2.7 did not allow for counting occurrences of a particular Licence Issue, nor did they allow quantification of the time extent of any such limitation. For these

reasons the binary flag system seemed to be the most feasible choice. The Licence Issues and the Voluntary Erasure flag were used as a measure of clinical performance and attitudes, while Erasure for Administrative Reasons was used as a measure of professional attitude (conscientiousness). The binary character of the LRMP criteria would result in weaker relationships with MRCP(UK) due to their limited variability.

#### **3.3.1.7 Investigation by the GMC - Fitness to Practice review**

Due to extreme sensitivity of the data and with respect to the fact that the majority of the information on the file was qualitative, only one criterion based on this data was employed: a binary variable where '1' was assigned to a doctor who was on the GMC FtP file, and '0' if they were not investigated by the GMC. The investigation outcomes or their reasons were not used.

#### **3.3.2 Variables**

Table 9 compiles the information on variables used in the analyses in this research. It indicates the source of criterial data, the criterion measures, the constructs that they represent, and the variables quantifying the criteria. Whenever more than one attempt in an exam was available, only the 1<sup>st</sup> attempt scores were employed. Some measures were assigned to both clinical skills and attitudes, as explained in Chapter 2. General performance measures were associated with all three components.

Also, four key demographic variables were employed as independent variables:

- Gender (1: female, 0: male),
- Ethnicity (1: white, 0: black or minority ethnic),
- Primary qualification (1: UK graduates, 0: International Medical Graduates)
- Probable UK trainee (1: UK trainee, 0: not a UK trainee) – this variable was set based on two premises: an MRCP(UK) candidate had to have a GMC number and a UK permanent address, which would suggest they were employed in the UK and they were in or were trying to get into CMT.

**Table 9. Key variables used in the analyses in this research with the represented construct, criterion source, and level of measurement.**

<i>Source</i>	<i>Criterion</i>	<i>Construct (measure of)</i>	<i>Variable</i>	<i>Level of measurement</i>
<b>MRCP(UK)</b>	Performance in Part I	knowledge	Mark on 1 <sup>st</sup> Attempt in Part I	ratio, continuous
	Performance in Part II	knowledge	Mark on 1 <sup>st</sup> Attempt in Part II	ratio, continuous
	Performance in PACES	clinical skills and attitudes	Mark on 1 <sup>st</sup> Attempt in PACES	ratio, continuous
	High performance in MRCP(UK)	general performance	MRCP(UK) Highfliers	ordinal (binary)
<b>MRCGP</b>	Performance in AKT	knowledge	AKT Overall Score (1 <sup>st</sup> attempt)	ratio, continuous
	Performance in Clinical Medicine	knowledge	AKT Clinical Medicine Score (1 <sup>st</sup> attempt)	ratio, continuous
	Performance in Evidence Interpretation	knowledge	AKT Evidence Interpretation Score (1 <sup>st</sup> attempt)	ratio, continuous
	Performance in Organisation	knowledge	AKT Organisational Questions Score (1 <sup>st</sup> attempt)	ratio, continuous
	Performance in CSA		CSA Equated Score (1 <sup>st</sup> attempt)	ratio, continuous
		clinical skills and attitudes	CSA Old Scheme Mark (1 <sup>st</sup> attempt) CSA New Scheme Mark (1 <sup>st</sup> attempt)	
<b>FRCR</b>	Performance in FRCR1	knowledge	Total Number of Attempts in FRCR1	interval
	Performance in FRCR1	knowledge	1 <sup>st</sup> Attempt Mean Module Mark from the Modules (FRCR1)	ratio, continuous
	Performance in Biology module	knowledge	Total Number of Attempts in Biology(FRCR1)	interval
	Performance in Clinical Pharmacology module	knowledge	Total Number of Attempts in Clinical Pharmacology (FRCR1)	interval
	Performance in Medical Statistics module	knowledge	Total number of Attempts in Medical Statistics (FRCR1)	interval
	Performance in Biology module	knowledge	Total Number of Attempts in Physics (FRCR1)	interval
	High Performance in FRFR1	knowledge	FRCR1 Rank	ordinal

**Table 9. Key variables used in the analyses in this research with the represented construct, criterion source, and level of measurement (continued)**

<i>Source</i>	<i>Criterion</i>	<i>Construct (measure of)</i>	<i>Variable</i>	<i>Level of measurement</i>
<b>FRCR</b>	Overall Number of Attempts in FRCR2	general performance	Total Number of Attempts in FRCR2	interval
	Performance in FRCR2	general performance	Pass/Fail Score in FRCR2	nominal (binary)
	Performance in the written test	knowledge	1 <sup>st</sup> attempt score in Written Exam (FRCR2)	ratio, continuous
	Performance in the clinical assessment	clinical skills and attitudes	1 <sup>st</sup> attempt score in Clinical Exam (FRCR2)	ratio, continuous
	Performance in the oral exam	knowledge	1 <sup>st</sup> attempt score in Oral Exam (FRCR2)	ratio, continuous
	High Performance in FRCR1	general performance	FRCR1 Rank	categorical
<b>SCEs</b>	SCE performance	knowledge	1 <sup>st</sup> Attempt score	ratio, continuous
<b>CKBA</b>	CKBA performance	knowledge	1 <sup>st</sup> Attempt score	ratio, continuous
<b>ARCP</b>	Overall Progress Assessment	clinical skills and attitudes	Satisfactory Outcomes Throughout Training	nominal (binary)
<b>LRMP</b>	Conscientiousness	attitude	Not registered due to administrative reasons	nominal (binary)
	Underperformance	clinical skills and attitudes	Licence Issues	nominal (binary)
	Underperformance	clinical skills and attitudes	Relinquished Licence (Voluntary Erasure)	nominal (binary)
<b>Fitness to Practice</b>	Underperformance	clinical skills and attitudes	Being on the list of investigated doctors	nominal (binary)

### **3.4 DESIGN**

This research employed a correlation design; however, based on its aims and the hypotheses all MRCP(UK) variables were treated as independent variables, while all performance measures from the described sources of data were considered dependent. The design also contained elements of retrospective longitudinal design (due to following-up on individual results) and cross-sectional design (when looking for differences between specified groups, based on e.g. FRCR1 Rank).

### **3.5 MISSING DATA**

There were several aspects of missing data – classified as both systematic and random – that required attention. All such issues are described in the following sections.

#### **3.5.1 Systematic Missing Data Issues**

The first systematic data issue was caused by the lack of electronic records of FRCR1 scores pre-2006. The missing data could not be recovered, and therefore, a variable based on the Total Number of Attempts in Modules of FRCR1 was created as a proxy measure of achievement in FRCR1 pre-2006. The variable has been described above.

The second systematic missing data issue was caused by time constraints, or cut-off points, for the datasets. The majority of the datasets used in this research, i.e. FRCR, MRCP(UK), MRCGP, LRMP, and ARCP, contained information that represented a continuous process not a singular event. Therefore, imposing time limits on datasets resulted in a partial lack of information on some of the individuals in those datasets, which is consistent with left- and right-censoring. A case of left-censoring occurred when some of the doctors did not have a full record on their performance prior to the start date of the dataset. The left-censoring effect was limited by providing information on the Total Number of Attempts. The right-censoring effect was observed when a candidate could not have been monitored throughout the examination process until a definite pass or fail. For example, it might have been that a candidate was still in the process of passing MRCP(UK) and e.g. their PACES result was missing. Another example comes from the FRCR1 dataset, where the Total Number of Attempts in Modules of FRCR1 was used. Some candidates might have still been in the two-year time-frame allowed for passing FRCR1. Therefore, their total number of attempts would be low, and if treated as final, it would indicate a higher ability than in reality. FRCR1 censored candidates were those who did not pass all of the modules at the closing date of the dataset, but their first recorded attempt was a maximum of three

sessions before the dataset closing date. Those candidates were excluded from inferential analyses. Censoring was present in the datasets of other examinations as well; but with the exception of the FRCR1 dataset, it did not have other effect on the analyses than limiting the sample sizes.

The third systematic missing data issue referred to international candidates for MRCP(UK). Doctors who do not practice in the UK are not required to obtain a GMC Reference Number. Many medical Royal Colleges allow their examinations to be taken by suitable candidates who can provide evidence of appropriate qualifications, which can be obtained anywhere in the world. Therefore, International Medical Graduates ('IMGs') who did not practice in the United Kingdom at the time of attempting any of the exams included in the analysis may not have been registered with the GMC. As the GMC Number was the primary matching key for the datasets, and other matching strategies were not always effective, there were some individuals whose data could not be matched and who were therefore excluded from the analysis. This could have potentially led to unjustified inferences when apart from criterial variables also nationality, ethnicity, PMQ, or UK training variables were taken into account. For that reason the analyses involving these variables were only seldom performed.

The fourth missing data issue resulted from the fact that exams are selection procedures, and as such they are affected by the range restriction problem, which is consistent with truncation of the data. Simply put, those who did not pass MRCP(UK) were not allowed to enter subsequent exams. This was the case with SCEs and equivalent exams. Within the MRCP(UK) those who did not pass Part I were not admitted to attempt Part II or PACES. The direct effect of selection is the decreasing variance in performance among the candidates (meaning that candidates become less varied in terms of their ability), and this limitation of variance can lead to underestimation of the true strength of the relationship between the analysed measures. Therefore, the validity coefficients based on exams were corrected for range restriction, as described in section 3.7.3.

### **3.5.2 Random missing data**

The random missing data mostly occurred within the demographic variables in the datasets. Because the data were missing at random, it was considered as not having an effect on the predictive validity inferences. The per cent of missing data varied depending on the dataset and the numbers are indicated in the descriptive statistics.

### **3.5.3 Handling missing data**

In general the analyses were performed on available data. While analysed, cases with data missing were deleted pairwise; missing values were not replaced using any of the existing algorithms. The Estimation-Maximization algorithm – hereinafter referred to as EM or EM algorithm – was, however, employed as a method of handling range restriction (Dempster, Laird, & Rubin, 1977). The EM algorithm is considered an effective alternative to the regular range restriction correction (Wiberg & Sundström, 2009), and therefore, it allows for estimating the true relationship between two measures. More on the method and application of the EM algorithm as correction for range restriction is presented in section 3.7 on statistical methods.

## **3.6 PROCEDURE**

The data provided for the purposes of this study were collected by independent entities during several assessment processes. Therefore, the datasets required reviewing, cleaning, merging, and anonymising. The review of each dataset was followed by data reorganisation and computation of new useful variables when required. For example, that was the case with the Total Number of Attempts in Modules of FRCR1, or obtaining Z-scores. The datasets were merged in a stepwise manner and involved pairs of datasets: the History File and at least one other dataset containing criterial data. Therefore, several separate files were created, one for each source of data. The merge was based on one of the two universal keys: the GMC Number or the RCP Candidate Number. The final files were anonymised through removal of identifiers allowing recognition of any individual doctor, such as the universal key, names, or date of birth.

The following list contains final set of datasets that were created and analysed:

- the History File, as described above in the materials section, dated April 2011
- the MRCP-FRCR file, dated May 2011, amended October 2012
- the MRCP-SCEs file (with Cardiology included), dated January 2012
- the MRCP-LRMP file, dated November 2012
- the MRCP-MRCGP file, dated January 2013
- the MRCP- ARCP file, dated March 2013
- the MRCP-GMC FtP panel file, dated May 2013

### **3.7 STATISTICAL TREATMENT**

The hypothesis for this research was that there exists a relationship between MRCP(UK) and subsequent measures of performance representing medical knowledge and skills and attitudes. All such relationships were analysed pair-wise between separate parts of MRCP(UK) and the selected measures of performance. The statistical treatment employed both descriptive and inferential univariate and multivariate statistics.

Calculation of the descriptive statistics served two purposes: enhanced understanding of the data and verification of assumptions for subsequent inferential statistics (for example, on the parameters of the distributions). The subsequent inferential statistics in each case served as a verification of existence of a statistical relationship. All statistical methods applied in the research are described in the following sections.

#### **3.7.1 Correlation coefficients**

Correlation coefficients were employed to establish the size of the general relationship between any two variables without assuming any causality between them. In the majority of cases parametric Pearson's  $r$  coefficient was calculated. In cases when the variables were continuous, but not normally distributed the relationships were quantified as Spearman's  $\rho$  or Kendall's  $\tau$  coefficients, or were approximated by bootstrapped Pearson's  $r$ . For categorical variables the Chi-squared statistics was provided with either Phi or Cramer's  $V$ . The magnitude of correlation coefficients was interpreted in accordance with Cohen's guidelines, based on which coefficients of approximately 0.10 are small, of 0.30 are medium, and of 0.50 are large (Hemphill, 2003); however, it needs to be remembered that these terms are relative.

#### **3.7.2 Correction for attenuation (disattenuation)**

Attenuation is an effect resulting from unreliability of the measurement tools (see Gulliksen, 1950). Correlation between two measures obtained with imperfect tools is burdened with certain error, proportional to the unreliability of both tools. In order to correct for that error the correlation coefficient is divided by the square root of the product of the two reliabilities. Disattenuation is required to establish the true size of the relationship between two measures. In the case of this research, it was applied simultaneously with the correction for range restriction.



### 3.7.3 Correction for range restriction

Range restriction is an effect of underestimating the size of a relationship, as a result of a selection procedure that limits the variance of scores in the final sample. A correction was required in this study as some of the criterion measures were exam scores that depended on passing MRCP(UK) first; for example SCEs or equivalent exams. Two methods of correction procedures were used depending on the measure corrected: one for the correlation coefficients, and another for the linear regression coefficients.

The correction for correlation coefficients was based on the ratio between standard deviation in an incumbent sample and the standard deviation in the unrestricted sample (Hunter, Schmidt, & Le, 2006; Hunter & Schmidt, 2004; Mendoza, 1987; Sackett & Yang, 2000). The procedure also required correction for attenuation of the used measures, but only if the reliability coefficients were known. Despite the fact that parts of MRCP(UK) constitute a sequential selection process, in this research they were treated as separate predictors, which allowed for use of the Thorndike (Case 2) method. The procedure applied was based on Stauffer & Mendoza (2001).

In the case of correction for the effect of range restriction on linear regression coefficients the EM algorithm (Dempster *et al.*, 1977) was employed, as it was proven by Wiberg & Sundström (2009) to be an effective alternative for more complicated methods. The method is based on the fact that a direct effect of range restriction is missing data in subsequent assessment for those candidates who failed to pass the initial one. Based on the available correlation matrix and distribution parameters of the results in both assessments the algorithm estimates the missing values and allows for calculation of the corrected statistics (in the case of this research regression coefficients). The SPSS 21.0 EM algorithm was used.

The applicability of the EM algorithm usually depends on the Missing Completely at Random assumption ('MCAR'). In concordance with this assumption there should be no relationship between the missing values and the values in the restricting exam. The assumption is tested with Little's MCAR test (Little, 1988). With regards to employing the EM algorithm as means for correcting for range restriction, Wiberg and Sundström (2009) assumed that the MCAR condition does not need to be met and instead, a weaker assumption of values Missing At Random ('MAR') sufficed, which is in line with Little (1992). MAR condition requires only that the probability of missing any value of the subsequent test was only dependent on the results of the pre-selection tool and not on the second test

itself. This interpretation was applicable to the relationships between MRCP(UK) parts and the criterion exams, and therefore, the values for Little's MCAR test were not necessary and were not provided.

#### **3.7.4 Comparison of means and analysis of variance**

A contrasting groups method allows for comparison of two or more groups based on a factor that results from an aggregation of various influences (Anastasi & Urbina, 1997, p.122). In other words, the comparison method facilitates a search for systematic changes or patterns of relations based on a specified factor, without that factor being a reason for those changes. For example, the comparison method allowed one to test whether the group of Highfliers in FRCR1 Rank was significantly better in MRCP(UK) parts than other candidates, regardless of the cause for their better performance. Depending on the number of compared groups either comparison of means (two groups) or analysis of variance (more than two groups) was employed whenever a categorical variable was a measure of performance, e.g. FRCR1 Rank or Satisfactory Outcomes Throughout Training, or LRMP status flags.

The comparisons of means were performed as independent samples t-tests and as Mann-Whitney U tests. For the purposes of meta-analyses the results of parametric comparisons were converted into effect sizes (point-biserial correlation coefficients) (Fritz, Morris, & Richler, 2012; Lyons, 1998).

#### **3.7.5 Linear regression**

Linear regression procedure allows for fitting the experimental data to a linear prediction function, which estimates the strength between the predictors and the dependent variable. Linear regression coefficients indicate direction and assume causality in a relationship. The linear models were applied in this study as means of quantifying the predictive effect of MRCP(UK) results on continuous criterion measures.

The assumption of no multicollinearity was made based on the average Variance Inflation Factor ('VIF'). This assumption is met when VIF falls below the critical value of ten (Field, 2009). Another assumption for verification was the independence of errors, which when not met can lead to an overestimation of the impact of the independent variables on the dependent variable. The assumption of independence of errors is usually recommended to be tested with the Durbin-Watson test (Durbin, Watson, & Durbin, 1950; Field, 2009). However, the Durbin-Watson procedure is mostly applicable for time-series data, which

was not the case for the analyses performed within this research. Hence, Durbin-Watson statistics were not provided across the results.

### 3.7.6 Chow test

The Chow test was employed for testing differences between *pairs* of linear models (Chow, 1960). The procedure described by Chow is based on a comparison of the residuals for any two models. The method assumes that if the models are represented by equations:

$$\text{Exam 1 Results} = \alpha_1 + \beta_1 \text{Part1FirstAttempt} + \gamma_1 \text{Part2FirstAttempt} + \lambda_1 \text{PACESFirstAttempt}$$

$$\text{Exam 2 Results} = \alpha_2 + \beta_2 \text{Part1FirstAttempt} + \gamma_2 \text{Part2FirstAttempt} + \lambda_2 \text{PACESFirstAttempt}$$

then they can be compared to a general model:

$$\text{Results} = \alpha + \beta \text{Part1FirstAttempt} + \gamma \text{Part2FirstAttempt} + \lambda \text{PACESFirstAttempt}$$

The method in fact assumes that  $\alpha_1=\alpha_2$ ,  $\beta_1=\beta_2$ ,  $\gamma_1=\gamma_2$ ,  $\lambda_1=\lambda_2$ . The Chow test compares the residuals after fitting the two compared models to residuals of the general model. The formula is expressed with the following F-test:

$$F(k, N1 + N2 - 2k) = \frac{\frac{SSR - (SSR1 + SSR2)}{k}}{\frac{SSR1 + SSR2}{N1 + N2 - 2k}}$$

Where:

$k$  is the number of parameters in the model including the intercept,

$N1$  and  $N2$  are the numbers of cases for respective models, and

$SSR$ ,  $SSR1$  and  $SSR2$  are the values of residual sum of squares for the respective models: general, model 1 and model 2.

The Chow test F-values were compared to critical values for an F -test with  $k$  and  $N1+N2-2*k$  degrees of freedom at a significance level of 5%.

### 3.7.7 Linear multi-level models

Linear regression multilevel modelling (Kreft & de Leeuw, 1998; Steenbergen & Jones, 2002) was employed to test if there were any differences between *more than two* fitted models, i.e. for the SCEs. Multi-level modelling requires nested or hierarchical data. In the case of SCEs nesting occurred because individual candidates were grouped into specialties. The model had two levels of variance: an individual level and a specialty level. The comparison is based on a similar assumption to the Chow test, i.e. that each SCE model can

be described using a general equation, where the parameters vary depending on the individual performance of candidates and on their specialty:

$$SCEfirstAttempt_{ij} = \beta_{0ij} + \beta_{1j}Part1FirstAttempt_{ij} + \beta_{2j}Part2FirstAttempt_{ij} + \beta_{3j}PACESFirstAttempt_{ij}$$

where:  $i$  indicates individual (1<sup>st</sup>) level,  $j$  indicates specialty (2<sup>nd</sup>) level

The study employed a random effects model, which means that analysed exams results were a sample of cases from a population, and therefore, the coefficients were estimated with an additional error (sampling error). The model for the SCEs was fitted using MLwiN v. 2.28 (Rasbash, Charlton, Browne, Healy, & Cameron, 2005). In order to centre the data, standardised first attempt results for all SCEs were used. Standardisation of each SCE result separately was equivalent to centring within clusters (Enders & Tofighi, 2007).

### 3.7.8 Logistic regression

Multiple logistic regression is based on fitting a model, where a categorical dependent variable is predicted by multiple predictors using a logistic function. The logistic function assigns probabilities (similar to coefficients in linear regression) to independent predictors that would lead to a dependent variable being 1 rather than 0 (Burns & Burns, 2009). That relationship is expressed either in the form of the log odds or in the form of the more convenient odds ratio, which represents the effect size. Odds ratios show a potential change in the dependent variable when a predictor changes by one unit. If the odds ratio is higher than '1' this means an increase of likelihood of occurrence of an event defined by the dependent variable, with every unit increase in the independent variable. For example, if passing ('1') or failing ('0') an exam is predicted by number of years of training with an odds ratio of 2, it means that with every additional year the odds of passing the exam double. If the odds ratio is smaller than '1' the effect is inverse – with every increase in the independent variable, the likelihood of occurrence of an event decreases. The odds ratio never falls beneath 0; however, the closer the ratio is to '0' the bigger the effect. The relative relevance of predictors in logistic regression models is assessed with the Wald statistic, which primarily is a statistical test for assessing if a particular predictor is likely to be 0 (null hypothesis for this test). The value of Wald statistic is calculated as the squared ratio of the coefficient divided by its standard error. The higher the Wald statistic is, the higher the impact of a particular variable on the model (Bewick, Cheek, & Ball, 2005).

Logistic regression was always employed when the dependent variables were binomial. For example, it was used to test if MRCP(UK) parts are predictive of: the Pass/Fail Score in FRCR2, the LRMP Licence Limitations, being under the investigation by the GMC, or the ARCP general outcome.

### **3.7.9 Meta-analysis**

Meta-analysis allows for an estimation of the average effect of a certain factor from a series of independent studies. In other words, it allows verification of a hypothesis of a general common effect being present in all studies included in the analyses. There are two types of the meta-analytical models: fixed-effect and random-effects (Borenstein, Hedges, Higgins, & Rothstein, 2009; Schmidt, Oh, & Hayes, 2009). Fixed-effect models assume that the true effect size underlying the investigated studies has a fixed value. Random-effects models assume that that effect size may differ depending on the sample on which it was tested; hence, there is more than just one effect size and the distribution of those effect sizes should vary around a mean with a certain deviation ( $\tau$ ). This means that in the random-effects models two sources of variance are accounted for: the variance of the effect sizes, referred to as the 'true' variability, and the sampling error. In the case of the present study, due to the impact of Part I, Part II and PACES on the analysed criterion measures potentially being different depending on the type of exam and its participants and conditions, the random-effects model was chosen. The Restricted Maximum Likelihood ('REML') estimation procedure was employed (Patterson & Thompson, 1971).

The meta-analytical model can be performed on any effect size measure such as Cohen's  $d$ , Pearson's  $r$ , and odds ratios. The models in this research were estimated on Pearson's  $r$  correlation coefficients corrected for artefacts, namely range restriction and attenuation (unreliability), or on mean difference effect sizes converted into point biserial correlation coefficients ( $r$ ). The conversion into point-biserial coefficient was based on methods described in Fritz *et al.* (2012).

The meta-analyses were performed with the 'metafor' package written for R. The method followed the guidelines of the author of the package (Viechtbauer, 2010). The analyses of correlation coefficients were not based on Fisher's  $r$  to Z transformations due to the fact that such transformed coefficients are believed to be upwardly biased (Hunter & Schmidt, 2004) and more difficult to interpret (Hedges, 1989). Further, employing 'metafor' allowed for a correction for bias resulting from sampling (Hedges, 1989) thanks to a built-in function. The package also provided several statistics for assessment of heterogeneity in

the analysed sample of studies. Those statistics were based on Higgins and colleagues' works (Higgins, Thompson, Deeks, & Altman, 2003; Higgins & Thompson, 2002).

In particular, apart from Cochran's Q test for existence of heterogeneity, metafor provided two key measures:  $H^2$  and  $I^2$ .  $H^2$  and  $H$  express the level of heterogeneity in the study;  $H$  can assume any value above 1, and if it equals 1 or less it suggests absolute homogeneity.  $I^2$  is a transformation of  $H$  and is "a proportion of total variation in the estimates of treatment effects that is due to heterogeneity between the studies" (Higgins & Thompson, 2002, p. 1552). Therefore, the higher this ratio is ( $I^2$  is presented as a percentage) the more of the variability in the effect estimate results from heterogeneity rather than sampling error. All of these measures were used in the interpretation of the results of the analyses.

The 'metafor' package also allows one to plot forest plots for visualisation of effects, funnel plots for assessment of asymmetry or in the case of this research of a potential bias in choosing the studies (Sterne, Sutton, Ioannidis, & Terrin, 2011), and Galbraith plots for assessment of heterogeneity (Galbraith, 1988; Galbraith, 1990). Galbraith plots were not employed.

### **3.7.10 Bootstrapping**

Bootstrapping is a well-established analytical method (Tooney & Duval, 1993) that allows for the use of parametric statistics when the underlying assumptions are not met, e.g. when distributions are different from normal. Bootstrapping is therefore a convenient alternative to non-parametric statistics. Non-parametric methods are considered less powerful and they often do not allow for testing interactions (Gaito, 1959); they are based on ranks rather than means, which is less intuitive and more difficult to interpret, and therefore less informative. However, the alternative view supports their use especially with larger samples (Hunter & May, 1993), or even points out that in specific cases they yield more statistical power than parametric tests (Tanizaki, 1997). Bootstrapping assumes that sampling from a sample resembles the process of sampling from the population. Therefore, through multiple re-sampling bootstrapping provides an estimate of the appropriate statistic with an associated error. This can be interpreted as an estimate of the effect size in the population (Tooney & Duval, 1993) after eliminating the sampling error. Hence, it is distribution free. The bootstrapping methods used during this research were applied with a native SPSS module based on 1,000 repetitions in each case of the bootstrapped statistics provided in this thesis. Unless indicated, the bootstrapped samples were not stratified.

### 3.7.11 Structural Equation Models

Structural Equation Modelling ('SEMo') is a statistical technique that allows for testing if a theoretical model of causal linear relationships between multiple variables fits the observed data (specifically, if it reproduces the variance-covariance matrix between variables) (Bacon, Lynd Bacon & Associates Ltd, & SPSS Inc, 1997; Bagozzi & Yi, 1988; Bielby & Hauser, 1977; Bollen, 1989; Hooper, Coughlan, & Mullen, 2008; Schumacker & Lomax, 2010). SEMo allows for testing multiple linear equations at the same time. The benefits of using it may also include estimating multicollinearity of variables (Jagpal, 1982) and unreliability of the data (Yang & Green, 2010). In the current research, SEMo was used to estimate the joint effect of several demographic factors on MRCP(UK) scores.

SEMo models fall under the same restrictions of statistical inference as any other statistical technique. The primary measures of goodness of fit of a model are:  $\chi^2$  statistic, Goodness-of-Fit Index, and Adjusted Goodness-of-Fit Index. Other significant indexes include: a root mean square of approximation statistic ('RMSEA'), and the Tucker-Lewis Index ('TLI'), which indicates if the proposed model is different from a null model (a model where all correlations are zero). A SEMo model is assumed to fit well to the data (reproduce the variance-covariance matrix) when  $\chi^2$  statistic is higher than 0.05, TLI is higher than 0.95, while RMSEA is below 0.06 or 0.08 depending on the literature (Bacon *et al.*, 1997; Schumacker & Lomax, 2010; Statistical Support University of Texas, 2001).

## 3.8 STATISTICAL SOFTWARE

The analyses for this study were performed using the following statistical software:

- SPSS 19.0 and 21.0 (IBM Corp., 2010) for majority of analyses, including AMOS module for structural equation models
- *R* v. 3.0.2 (*R* Development Core Team, 2005) including 'metafor' package (Viechtbauer, 2013) for meta-analyses
- MLwiN v. 2.28 and 2.30 (Rasbash *et al.*, 2005) for the purposes of multilevel modelling
- Microsoft Excel (2007) for minor computations on correction for attenuation and range restriction
- Matlab (The MathWorks Inc., 2010) was used by Chris McManus to write a programme that would read the fifty monthly LRMP files, as mentioned in section 3.3.1.6 on measures obtained from the LRMP.

### 3.9 LIMITATIONS

This research was designed purposefully as a qualitative study with an acknowledgment of certain limitations arising from this approach and its methods.

A typical predictive validity study links results of an exam with a criterial measure observed after a certain period of time. Therefore, a predictive validity study by default assumes a longitudinal methodology. However, a purely longitudinal study would require decades of data, in which case it should have been planned ahead many years ago. Alternatively, a longitudinal study of a three-year time-frame typical for a PhD programme may not have been sufficient to yield any conclusions. In order to resolve this apparent conflict, a retrospective longitudinal correlational approach on historical MRCP(UK) data was chosen, which allowed the extension of the time-frame of the study up to eight years.

It was a subjective but conscious choice driven largely by feasibility and a short time-frame of the project that this research would not be based on collection of data by survey or qualitative methods, but would be based solely on analyses of existing data. The first method would involve addressing MRCP(UK) candidates and physicians directly with a questionnaire, which would be a costly and time-consuming way of gathering data. Further, the questionnaires would contain questions about doctors' current practise, previous certificates, personal characteristics, or some specific chosen measures of performance, such as results of assessments, number of experienced complaints, or review proceedings, some of which would be of delicate nature and could be perceived as threatening. Such a survey would, therefore, provide the researcher with self-reported data, which are susceptible to biases. For example, the need for social approval or social desirability can seriously skew results (Crowne & Marlowe, 1964; DeMaio, 1984). It was also taken into account that the results of surveys could be biased depending on the method of collection. An on-line survey would probably have a low response rate which would result in a small sample size, and a paper form would be difficult and expensive to convey as there are approximately 6,300 MRCP(UK) candidates every year from around the world. Therefore, the completeness of the survey response data was a serious risk for this project. It would extend the time-frame of the research or otherwise it could create a systematic missing data issue and lack of reliability that would require further investigation. For similar reasons, namely time- and cost-efficiency, with additional aspect of even smaller sample sizes it was decided that this research would not contain a qualitative component such as observations or interviews.



The choice of obtaining and analysing the existing data alleviated most of the above-mentioned problems, but was associated with other limitations. Firstly, only quantifiable measures were chosen. This meant that descriptive data, such as opinions, reports, evaluations, etc., were not considered as feasible sources of criteria, despite their potential value. In the case of measures such as registration status changes or ARCP assessments a qualitative approach could have provided additional insight into the processes or individual decisions, but for practical reasons it was decided not to be undertaken. Secondly, adopting a retrospective approach meant that the criterion measures were not purposefully designed for the study, but rather stemmed from what was attainable, which could be associated with a 'data-dredging' bias. Further, employing only pre-existing data in the worst-case scenario could have resulted in not finding any suitable criteria for any of the three components of competence due to a limited scope of available measures. In fact, this difficulty was already discussed in the case of the attitude component. The binary interpretation of the ARCP outcomes and further aggregation of these classifications into one dichotomous measure necessarily resulted in decreased variability of the analysed data, which would have an effect on the magnitude of the correlation coefficients between MRCP(UK) and ARCP. An alternative would have been to create an ordinal scale as done by Tiffin and colleagues (Tiffin, Illing, Kasim, & McLachlan, 2014). Finally, not all institutions administering the relevant data were inclined to provide access to their archives.

Despite those deficiencies the study based on pre-existing data seemed to be the most feasible approach of exploring the subject of the predictive validity of MRCP(UK). The potential sources of criterion measures, their selection and methodology have already been thoroughly described.

### **3.10 ETHICAL APPROVAL**

The present studies were exempt from UCL Research Ethics Committee approval, as they involved analyses of anonymised pre-existing data consisting of results of educational tests and assessments. This practice is consistent with the UCL Research Ethics Committee Resources available on-line (UCL Research Ethics Committee, 2011). This has been confirmed by the (then) Chair of the UCL Research Ethics Committee, Sir John Birch, in his e-mail dated February 15<sup>th</sup> 2011. Contents of the enquiry and Sir John Birch's response constitute Appendix A.

## SUMMARY

This research adopted a correlational design with elements of both longitudinal and cross-sectional approaches and was based solely on pre-existing data collected by various entities involved in the medical education system. As such, this study was exempt from the UCL Ethics Committee Approval. The reasoning for choosing this particular design was that analysis of the retrospective data should be cost- and time-efficient in comparison to a survey study, and that it would provide more objective results particularly when the criteria may suggest deficiencies in medical proficiency. However, it was acknowledged that a purely quantitative approach also have certain deficiencies, especially in case of criteria that are contextual, such as Licence Issues or ARCP assessment outcomes.

Due to the chosen method of conducting this research several methodological aspects needed to be addressed. Among others this study was affected by missing data issues, including left- and right-censoring, and data truncation. It was also susceptible to any difficulties in the data merging procedure, as the sample sizes depended largely on the missing universal key variables such as the GMC Number or the RCP Candidate Number, which could not have been helped or avoided. This might have had an effect when ethnicity or PMQ effects were investigated, as majority of missing cases were IMGs. Missing information on, for example, gender was random and therefore should not bias the inferences made based on analyses.

The largest part of this chapter was devoted to the materials, which were the datasets secured for the purposes of this study, and the list of criterion measures and variables that quantified them. This description was provided in order to introduce the data that were to be analysed. The delineation of chosen statistical methods was included to introduce them and to address any questions that might have arisen with relation to their use. The actual results of the analyses constitute the following Chapters 4 to 7.

## Chapter 4

### The MRCP(UK) internal prediction results

#### ABSTRACT

*It was hypothesized that MRCP(UK) would be predictive of subsequent measures of performance in terms of knowledge and clinical skills. The MRCP(UK) parts could be treated as separate exams, and discovering the nature of relationships between them could serve two main purposes. First, it would support the validity of the MRCP(UK) exam as a whole. Second, the strength of the associations between them could provide arguments for aggregating partial results into one score for the purposes of subsequent analyses. Both aspects were considered. The results supported the validity of MRCP(UK); however, the obtained results did not provide sufficient evidence for employing one aggregate score. Additional analyses indicated that first attempt results were predictive of all subsequent attempts in any MRCP(UK) part, which allowed for employing first attempt results as the key measures of performance in the next steps of this research. The validity of MRCP(UK) was also investigated in terms of the effect of demographic characteristics of candidates on exam performance. It was found that gender, ethnicity, and UK primary medical qualification ('PMQ') and UK training had a significant effect on performance in MRCP(UK); however, UK PMQ played a predominant role among these factors. The obtained results were summarized and discussed.*

The analyses of the relationships between MRCP(UK) and external criteria were preceded by an analysis of the internal relationships between parts of MRCP(UK). This analysis was performed for two major reasons. Firstly, MRCP(UK) may take up to several years to complete, and therefore, its parts can be treated as separate exams. Referring to the previously addressed concept of the Academic Backbone (McManus *et al.*, 2013) and the accumulation of the knowledge and skills in medicine, the separate parts of MRCP(UK) were likely to predict one another. In Chapter 1, it was hypothesised that MRCP(UK), in general, would predict future knowledge tests and measures of clinical performance, i.e. skills and attitudes. Therefore, by extension, it was likely that Part I scores would predict scores in Part II and PACES and that scores in Part II would predict those in PACES. The existence of such relationships would support the predictive validity of MRCP(UK), while at the same time supporting the purposefulness of its design. Secondly, if the results from different

MRCP(UK) parts correlated highly, MRCP(UK) parts results could be aggregated into one measure which would affect the approach towards further analyses. Therefore, the pattern of the internal relationships was studied to address the above.

## 4.1 DESCRIPTIVE STATISTICS

The analyses were performed on the main History File. It contained 50,311 records from exam diets taking place within a period of eight years, from May 2003 to January 2011. Therefore, the annual average number of candidates was 6,288. Based on the contents of the file, 58.3% of the candidates were male (29,326), and 41.7% were female (20,971). Of those candidates who disclosed their ethnicity 26.1% (13,146) declared themselves white, and 73.9% (37,165) declared themselves non-white. Of the total number of records, 38.4% (19,342) were those of UK graduates, and approximately 56.8% (28,580 individuals) were those of UK trainees. UK training was not self-declared; it was established based on the existence of the GMC Number (which means that a doctor can practise in the UK), and on the presence of a current UK correspondence address.

The mean time interval between the first attempt at Part I and passing PACES was approximately 131 weeks; almost 2.5 years. Figure 8 (below) presents a timeline (in weeks ) of the key stages in the process of passing MRCP(UK), including the first attempts in all of the parts.



*Figure 8. Timeline of key stages (mean number of weeks) of the MRCP(UK) process, from PMQ to passing MRCP(UK). Number of valid cases depended on the stage (due to censoring) from  $n=11,990$  to  $n=35,950$  candidates.*

Inspection of Figure 8 shows that the time interval between the dates of the first attempt and ultimate pass in Part I is shorter (ten weeks) than in the case of Part II and PACES (twenty and thirty weeks, respectively). These differences, however, should not be simply interpreted as Part II or PACES being more difficult, as this might have been the result of different delays between the diets for each of the parts. Table 10 shows the distribution

characteristics for the number of attempts and the number of attempts at which MRCP(UK) part was passed.

As indicated in the table the means for the number of attempts and the number of attempts at which a part was passed were similar (1.99 *versus* 1.96 for Part I, 1.70 *versus* 1.63 for Part II, and 1.90 *versus* 1.76 for PACES). The distributions of the number of attempts in all parts were significantly different from normal, as indicated by the statistics for skewness and kurtosis, and were confirmed by the results of the Kolmogorov-Smirnov tests ( $Z_{KS}$ ) also presented in Table 10.

It was revealed that 39,335 candidates had a record of having attempted Part I. The majority (24,586) eventually passed Part I, which was approximately 63% of the whole sample. More than a half of those who passed Part I (36% of the total sample, 14,127 candidates) did so on their first attempt. From an overall 23,637 candidates having attempted Part II, 89.9% (21,260) eventually passed the exam, and 61.4% (14,512) of the total did so on their first attempt. Of those who attempted PACES in the eight-year period (which was 21,270 candidates), 82.9% (17,629) eventually passed the exam, with a large group of candidates doing so on their first (48.67%; 10,353) or second (18.9%; 4,014) attempt. Detailed cross-tables presenting the number of passed and failed candidates depending on the total number of attempts are included in Appendix D. Among candidates who had a record of all attempts there were 4,025 individuals who passed the three parts of MRCP(UK) on their first attempt, which constituted 8% of the sample. Such candidates are referred to as the 'MRCP(UK) Highfliers'.

**Table 10. Distribution parameters for the number of attempts in MRCP(UK) parts variables.**

<i>Variable</i>	<i>N</i>	<i>Min</i>	<i>Max</i>	<i>Mean</i>	<i>SD</i>	<i>Skewness</i>	<i>Kurtosis</i>	<i>K-S test</i>
Number of attempts in Part I	39,335	1.00	26.00	1.99	1.74	3.19	15.96	$Z_{KS}=57.5, p<0.001$
Number of attempts in Part II	23,637	1.00	21.00	1.70	1.41	3.45	17.77	$Z_{KS}=54.5, p<0.001$
Number of attempts in PACES	21,270	1.00	14.00	1.90	1.37	2.28	6.96	$Z_{KS}=42.5, p<0.001$
Attempt Part I passed	24,586	1.00	26.00	1.96	1.66	3.12	15.79	$Z_{KS}=46.4, p<0.001$
Attempt Part II passed	21,260	1.00	19.00	1.63	1.30	3.41	16.76	$Z_{KS}=54.1, p<0.001$
Attempt PACES passed	17,629	1.00	12.00	1.76	1.22	2.33	7.24	$Z_{KS}=42.5, p<0.001$

The above data on attempts are affected by censoring, as mentioned in the Methodology (Chapter 3). This is reflected in the different numbers of valid cases for each of the MRCP(UK) parts. Anyone whose last recorded result was a fail could have been considered a drop-out, i.e. a person who resigned from passing MRCP(UK) entirely, or could have been still in the process of preparing for the next attempt (in which case the data would be right-censored). It was impossible to differentiate between those who dropped-out or were still preparing for their next attempt in MRCP(UK).

## 4.2 MRCP(UK) AS A PROCESS

The data discussed above presented a cross-sectional perspective, which was based on all available records in the dataset. However, those numbers did not indicate how truly selective MRCP(UK) is, and therefore, an alternative longitudinal approach was used to show MRCP(UK) as a process. This was based on those candidates who had a record of Part I and could have been followed through the process until passing PACES. Figure 9 (below) presents such a process in the form of a ‘funnel’ chart, based on the eight-year timeframe.

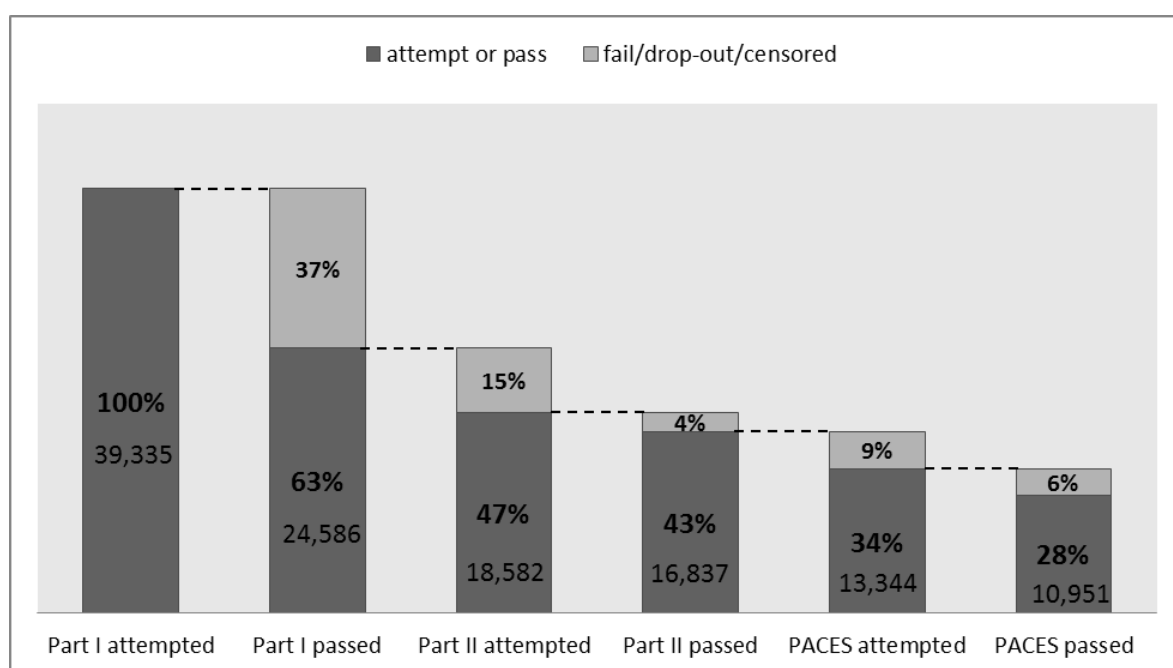


Figure 9. MRCP(UK) as a selective process: numbers of candidates after each stage of attempting the exam (based on  $n=39,335$ ).

Initially, there were 39,335 individuals (first bar on the left on Figure 9 marked as ‘Part I attempted’) who were known to have at least one attempt in Part I. This number was the base value for calculating the percentages in consecutive steps. Of this initial group of

doctors, 63% (24,586) eventually passed Part I. Of these 63%, almost three quarters attempted Part II (18,582 which is equal to 47% of the original number of cases) and the majority of them passed it (16,837 in total, constituting 91.5% of the number of cases from previous step; 43% of the initial number of candidates). And of these 43%, 79% (13,344) attempted PACES, which was equal to 34% of the original number of candidates. In the final stage, almost 28% (10,951; 44% of those who passed Part I) of the initial number of candidates completed MRCP(UK). As presented in Figure 9, the number of failed candidates includes those who were censored and still could have attempted any of the parts, although that becomes progressively less likely as the time intervals increase.

The effect of multiple attempts on MRCP(UK) results was studied from a different angle by considering educational progress of the candidates, which raised new questions: first, if doctors improve with each attempt; second, if it is possible that some candidates will never achieve the required level of ability however many attempts they have; and third, if those with insufficient level of knowledge or skills are likely to pass by chance.

### **4.3 LEARNING CURVES AND THE MEANING OF THE FIRST ATTEMPT SCORES**

Figure 10 shows the gradual increase in MRCP(UK) scores for Part I on consecutive attempts in several groups of candidates. Those groups marked with different colours are composed of candidates with the same overall number of attempts in Part I. For example, Group 5 consisted of candidates who had five attempts in Part I by January 2011, irrespective of the final outcome. Each line connects the marker points being a particular group's mean score on each consecutive attempt. For clarity of the graph, only the first three groups have the mean score values presented on the graph. The means were calculated based on the scores standardised against the pass-mark, meaning that '0' was the pass-mark, as in Figure 10.

The data of Figure 10 show that the higher the number of attempts overall, the lower the mean first attempt score. The actual means are provided in Table D4 in Appendix D.



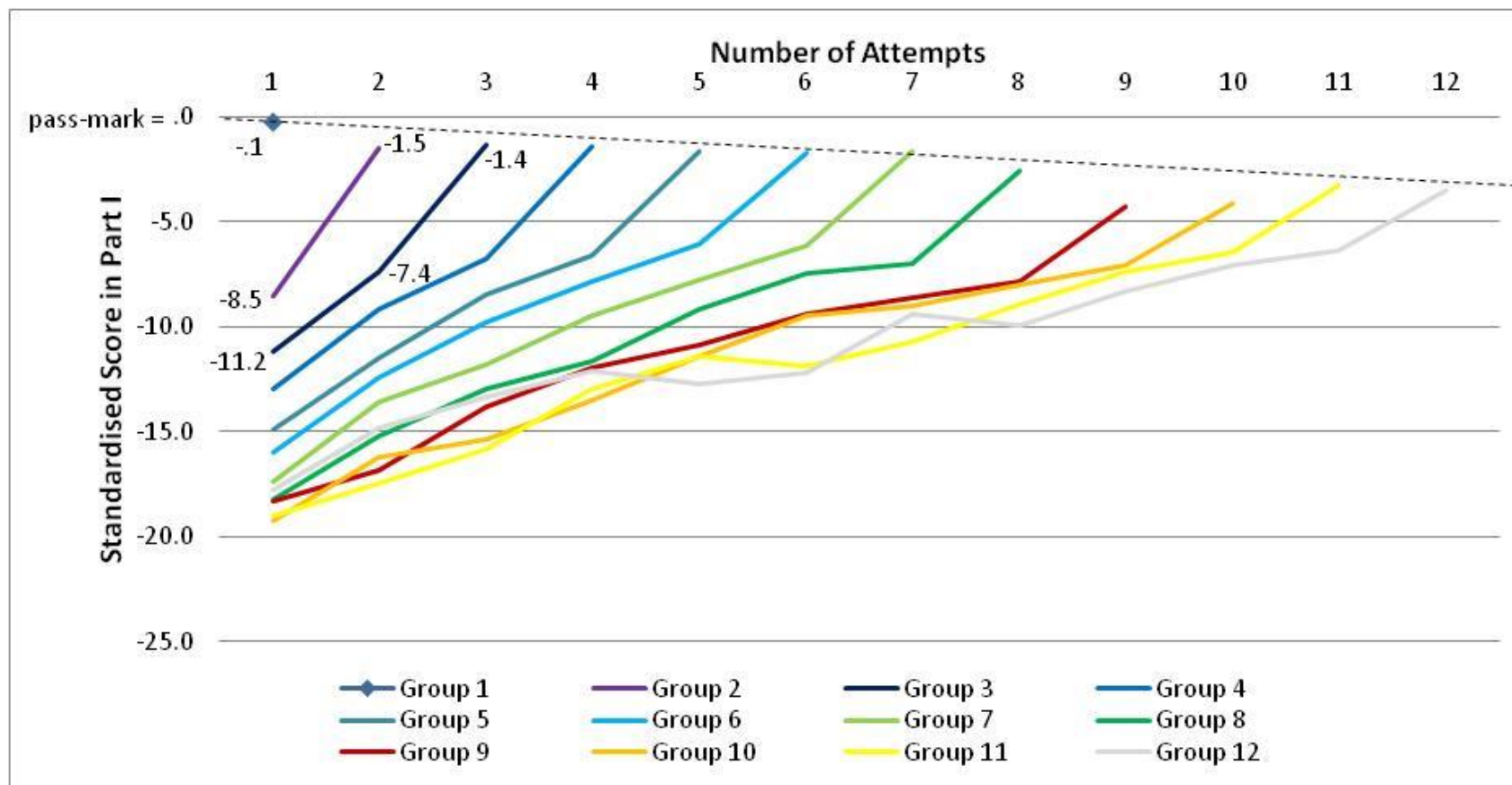


Figure 10. Mean scores in Part I per attempt for groups based on total number of attempts – approximation of the learning curve

The differences in Part I first attempt results between the twelve groups shown in Figure 10 were tested using bootstrapped one-way ANOVA due to the fact that the distributions were significantly different from normal (see Table 12), and the assumption of homogeneity of variance was not met ( $F(11, 35,950) = 287.27, p < 0.001$ ). The results of the test showed significant differences between the above-mentioned groups  $F(11, 35,950) = 801.24, p < 0.001$ . These results were corroborated by the Kruskal-Wallis H test, which was also highly significant ( $\chi^2(11, n=35,938) = 8,631.43, p < 0.001$ ). The similarities and differences between groups were shown using the Ryan-Einot-Gabriel-Welsch ('REGW') Q post-hoc test: candidates from groups 1 to 4 constituted separate homogenous categories, each with  $p=1.0$ . Candidates from groups 5 to 12 were homogenous at  $p=0.239$ . The results of the ANOVA and REGW Q tests for Part II and PACES are presented in Tables D7 and D8 in Appendix D to this thesis.

Further, with each consecutive attempt, there was a gradual increase in the average score, suggesting improvement in candidates' ability. Similar observations were made for Part II and PACES results. Due to the repetitiveness of those findings, graphs for Part II and PACES are only included in Appendix D, along with the tables of means for these parts (see Tables D5 and D6). The observed effect may have resulted from actual improvement in the tested ability; however, the educational literature also points out factors that may lead to better performance on resits that are irrelevant to the measured construct, such as practice (Reeve & Lam, 2005), or situational effects (Hausknecht, Trevor, & Farr, 2002; Matton, Vautier, & Raufaste, 2009). McManus (1992) and Hausknecht *et al.* (2002) did find however, that up to a certain point true ability does improve and an increase in scores cannot be associated purely with construct irrelevant factors. This was further investigated by McManus & Ludka (2012), who hypothesised that the performance in each group of candidates can be modelled with non-linear learning curves. Fitting such models could also provide estimates of how the candidates would have performed if they were allowed to take the MRCP(UK) parts after already passing them, which would be consistent with overcoming the problem of data truncation. The process of fitting the learning models was described in a paper on re-sitting high-stakes exams on the basis of MRCP(UK) data. Firstly, the curves were approximated by linear relationships between any two attempts using MLwiN (Rasbash *et al.*, 2005). The second phase required non-linear modelling using SAS (SAS Institute Inc., 2004). The models have shown that the gradual improvements of the MRCP(UK) candidates fitted the negative exponential curve typical for learning.

The key findings from that analysis were as follows:

1. there is a maximum level of achievable ability for each candidate, and that such level is an individual characteristic; in case of the candidates for whom that level falls beneath the pass-mark, the chance of passing the exam is negligible;
2. candidates actually improve in the tested ability rather than just obtain better scores due to luck ;
3. first attempt results are predictive of subsequent attempts results, as they are predictive of the candidate's improvement rate;
4. first attempt results in Part II and PACES correlate highly with Part I first attempt results.

These findings imply that first attempt results can be used as the key performance measures for inferential analyses. By extension, first attempt results in other criterion measures are also assumed to follow this pattern. In further analyses, therefore, this thesis mostly reports marks at first attempts, and not at subsequent attempts.

Inspection of Figure 10 also indicates that the last observed average result in each group never reached the pass-mark, which suggests that more candidates failed the exam rather than passed it, or that the scores of those who passed did not compensate for the results of those who failed. The dashed line shows a trend in decreasing final score means. Although it seems that the deviation from the pass-mark is much bigger in groups with higher numbers of attempts (groups 8 to 12) in comparison to the lower-number groups, that effect was probably due to a much smaller number of valid cases in the high-attempt groups. Table 11 summarizes the valid cases, mean scores at the final attempt in each group, and provides the standard errors of means for that last attempt score and 95%CI for the mean. It also shows that indeed the standard error increases with the group number and the confidence intervals for the means become wider.

**Table 11. Mean score at final attempt in Part I for groups distinguished based on the total number of attempts in Part I with SE, 95% confidence intervals, and numbers of valid cases.**

<i>Group</i>	<i>Mean</i>	<i>SE</i>	<i>95%CI</i>		<i>N</i>
			<i>Lower</i>	<i>Upper</i>	
group 1	-0.14	0.08	[-0.30,	0.02]	22,602
group 2	-1.46	0.11	[-1.67,	-1.24]	7,830
group 3	-1.36	0.14	[-1.64,	-1.07]	3,784
group 4	-1.39	0.18	[-1.74,	-1.04]	2,098
group 5	-1.63	0.24	[-2.10,	-1.17]	1,224
group 6	-1.73	0.30	[-2.33,	-1.14]	667
group 7	-1.65	0.38	[-2.40,	-0.91]	395
group 8	-2.60	0.49	[-3.57,	-1.63]	254
group 9	-4.26	0.65	[-5.54,	-2.98]	163
group 10	-4.13	0.78	[-5.66,	-2.60]	100
group 11	-3.27	1.15	[-5.52,	-1.02]	59
group 12	-3.50	1.07	[-5.60,	-1.40]	48

Table 12 presents descriptive statistics for the first attempt scores in MRCP(UK) parts. Despite the fact that the values of skewness and kurtosis suggest normality of the distributions, the Kolmogorov-Smirnov tests ( $Z_{KS}$ ) indicate clearly that the distributions were significantly different from normal.

The non-normality of the distribution would usually call for the use of non-parametric statistics; however, parametric methods are considered better in terms of power and robustness and they would provide better assessment in terms of multivariate analyses, as already argued in the Methodology chapter (section 3.7.10). Further analyses within this chapter were, therefore, performed using parametric, bootstrapped parametric, and non-parametric methods, in order to ascertain if use of parametric methods under lack of normality of the distributions would yield a significant error and lead to misinterpretation of the findings.

**Table 12. Distribution parameters for the first attempt results in the parts of MRCP(UK).**

<i>Variable</i>	<i>N</i>	<i>Min</i>	<i>Max</i>	<i>Median</i>	<i>Mean</i>	<i>SD</i>	<i>Skewness</i>	<i>Kurtosis</i>	<i>K-S test</i>
Part I 1st Attempt	35,962	-65.22	46.57	-3.50	-4.03	11.97	-0.21	-0.22	$Z_{KS}=5.55$ , $p<0.001$
Part II 1st Attempt	22,398	-30.92	42.85	3.29	3.27	7.628	0.06	0.55	$Z_{KS}=2.31$ , $p<0.001$
PACES 1st Attempt	21,260	-30.08	15.00	-0.90	-1.20	6.93	-0.43	-0.12	$Z_{KS}=9.05$ , $p<0.001$

## 4.4 CORRELATIONS BETWEEN FIRST ATTEMPT SCORES AT MRCP(UK) PART I, PART II AND PACES

### 4.4.1 Between First Attempt Scores at MRCP(UK) Part I, Part II and PACES

As MRCP(UK) is a three-stage exam and Part II and PACES can only be taken after Part I is passed, there is a time window between consecutive parts as shown previously in Figure 8. Therefore, MRCP(UK) parts could have been treated as separate exams and separate predictors rather than just one aggregate measure. In order to discern whether to treat MRCP(UK) parts as separate predictors or to aggregate their results into one, an assessment of the strength of the relationships between the MRCP(UK) parts was required.

The non-parametric Spearman's  $\rho$  coefficients, Pearson's  $r$ , and bootstrapped Pearson's  $r$  coefficients were calculated between first attempt scores in Part I, Part II and PACES, and the results are presented in Table 13.

**Table 13. Correlation coefficients (Spearman's  $\rho$ , Pearson's  $r$  and bootstrapped Pearson's  $r$ ) for first attempt scores at MRCP(UK) parts.**

<i>Relationship between</i>	<i>Spearman's <math>\rho</math></i>	<i>Pearson's <math>r</math></i>	<i>Bootstrapped Pearson's <math>r</math></i>
Part I and Part II	0.60**	0.61**	0.60** (bias=0.00, SE=0.005) 95%CI [0.60, 0.62]  <i>n</i> =16,744
Part II and PACES	0.38**	0.38**	0.38** (bias=0.00, SE=0.007) 95%CI [0.37, 0.40]  <i>n</i> =11,998
Part I and PACES	0.30**	0.30**	0.30** (bias=0.00, SE=0.008) 95%CI [0.28, 0.32]  <i>n</i> =16,561

\*\* significant at  $p < 0.001$

Due to the fact that MRCP(UK) is a sequential selection procedure, in which Part I selects those who are allowed to attempt Part II and PACES, and in which Part II limited access to PACES (until the end of 2008), the estimated correlation coefficients were affected by range restriction. In order to estimate the true strength of the relationships between MRCP(UK) parts, a correction was applied using the classic Thorndike (Case 2) method. Table 14 summarizes data that were required for the process of correction. The unrestricted sample standard deviation is the deviation of the first attempt scores for all candidates who had a

record of that exam in the History File. The restricted sample SD is the standard deviation of only those candidates who have both the record of the first and the subsequent exam. It was not taken into account that PACES is restricted not only by Part I, but also by Part II.

The correction for range restriction and disattenuation resulted in an increase in the correlation coefficients: between Part I and Part II it reached  $r=0.78$ ; between Part II and PACES  $r=0.48$ , and between Part I and PACES  $r=0.43$ . The obtained values differed slightly from those presented in the McManus and Ludka paper (McManus & Ludka, 2012), where the estimates of the true relationships in the sense of Structural Equation Modelling ('SEMo') were provided.

**Table 14. Correlation coefficients between MRCP(UK) parts with parameters required in the process of range derestriction and disattenuation**

<i>Relationship between</i>	<i>Bootstrapped Pearson's r</i>	<i>Reliability of the selecting exam*</i>	<i>Reliability of the consecutive exam*</i>	<i>SD restricted sample</i>	<i>SD unrestricted sample</i>	<i>Corrected coefficient</i>
Part I and Part II	0.60	Part I – 0.91	Part II – 0.81	9.52	11.98	0.78
Part I and PACES	0.30	Part I – 0.91	PACES – 0.82	9.25	11.98	0.43
Part II and PACES	0.38	Part II – 0.81	PACES – 0.82	6.86	7.61	0.48

\*Mean reliability of the exam between years 2003 and 2011.



#### 4.4.2 Between First Attempt Results and Overall Number of Attempts

The relationship between the first attempt results and the number of attempts in each of the MRCP(UK) parts was also calculated using three methods: parametric Pearson's  $r$ , bootstrapped Pearson's  $r$  coefficients, and for verification, also non-parametric Spearman's  $\rho$  coefficients due to non-normality of the distributions. The coefficients were calculated for the following pairs of variables: the First Attempt score in Part I and the Total Number of Attempts in Part I, the First Attempt score in Part II and the Total Number of Attempts in Part II, and the First Attempt score in PACES and the Total Number of Attempts in PACES. The coefficients are presented in Table 15.

**Table 15. Correlation coefficients (Spearman's  $\rho$ , Pearson's  $r$  and bootstrapped Pearson's  $r$ ) between first attempt scores and total number of attempts in MRCP(UK) parts – a comparison.**

<i>Relationship between</i>	<i>Spearman's <math>\rho</math></i>	<i>Pearson's <math>r</math></i>	<i>Bootstrapped Pearson's <math>r</math></i>
Part I 1 <sup>st</sup> attempt score and Total No of Attempts	-0.49**	-0.39**	-0.39**(bias =0.00, SE=0.004) 95%CI [-0.40, -0.38]  n=35,962
Part II 1 <sup>st</sup> attempt score and Total No of Attempts	-0.67**	-0.54**	-0.54**(bias=0.00, SE=0.005) 95%CI [-0.55, -0.53]  n=22,398
PACES 1 <sup>st</sup> attempt score and Total No of Attempts	-0.75**	-0.58**	-0.58**(bias=0.00, SE=0.006) 95%CI [-0.59, -0.57]  n=21,260

The negative sign of the coefficients indicates that with an increase in any of the first attempt results, the total number of attempts decreased. The magnitudes of the Spearman's  $\rho$  coefficient and Pearson's  $r$  coefficient for each pair of variables were quite similar and in accordance with Cohen's guidelines – large. Bootstrapped results indicate no bias.

## 4.5 CONTRASTING GROUPS

### 4.5.1 MRCP(UK) Highfliers

As was mentioned above, approximately 8.0% ( $n=4,025$ ) of candidates passed all MRCP(UK) parts on their first attempt. A univariate non-parametric comparison of means was performed and a bootstrapped independent samples t-test was employed to test whether

there were significant differences between MRCP(UK) Highfliers scores and the scores of typical candidates. The results are summarised in Table 16 (next page).

The analyses summarised in Table 16 showed that MRCP(UK) Highfliers indeed scored significantly higher than the rest of the candidates, and the effect sizes were strong. Clearly, the non-parametric and parametric tests led to the same interpretation, and bootstrapping confirmed the significance of the observed differences. To visualise the distance between the scores in both groups, Figure 11 presents the bootstrapped mean scores (with standard error of bootstrap estimation) for the Highflier and Typical Candidates groups for each part separately.

Inspection of Figure 11 suggests that the largest absolute value between the results of the two groups was observed for Part I scores. This is consistent with the assumption that Part I is the screening exam.

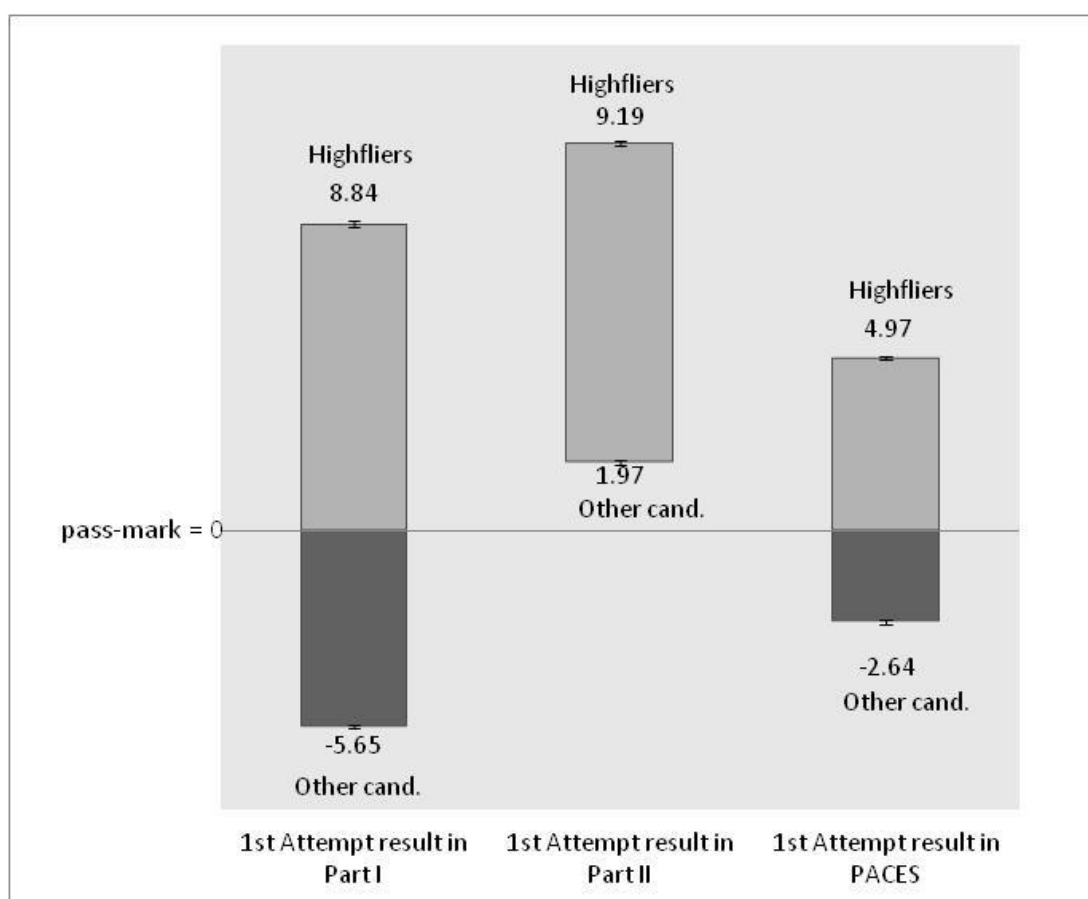


Figure 11. Visualised mean bootstrapped results (with standard error) in MRCP(UK) parts for the Typical Candidates and Highfliers groups (number of valid cases provided in Table 16).

**Table 16. Comparison of mean scores between the Highfliers and Typical Candidates groups in MRCP(UK) parts (independent samples t-test results with bootstrapping and Mann-Whitney Z results) with effect sizes.**

<i>Groups</i>	<i>N</i>	<i>Means and SDs</i>	<i>Bootstrapped Means and SDs</i>	<i>95% CI for the mean</i>	<i>Mann-Whitney Z</i>	<i>Independent samples t-test</i>	<i>Effect size (r)</i>
<b><u>Part I</u></b>							
Highfliers	4,025	8.84 SD=5.62	8.84 (bias=~-0.00, SE=0.088) SD=5.62 (bias=-0.003, SE=0.062)	[8.67, 9.01]	Z=-76.92, p<0.001	t(35,960)=-78.32, p<0.001	0.38
Typical	31,937	-5.65 SD=11.57	-5.65 (bias=-0.001, SE=0.066) SD=11.57 (bias=-0.002, SE=0.046)	[-5.79, -5.52]			
<b><u>Part II</u></b>							
Highfliers	4,025	9.19 SD=5.36	9.19 (bias =-0.002, SE=0.084) SD=5.36 (bias=~-0.00, SE=0.065)	[9.02, 9.35]	Z=-58.04, p<0.001	t(22,396)=-58.50, p<0.001	0.36
Typical	18,373	1.97 SD=7.42	1.97 (bias =-0.002, SE=0.054) SD=7.42 (bias=0.001, SE=0.047)	[1.85, 2.08]			
<b><u>PACES</u></b>							
Highfliers	4,025	4.97 SD=3.39	4.97 (bias =-0.004, SE=0.053) SD=3.39 (bias=0.003, SE=0.029)	[4.86, 5.08]	Z=-67.24, p<0.001	t(21,258)= -69.48, p<0.001	0.43
Typical	17,235	-2.64 SD=6.75	-2.64 (bias =~-0.00, SE=0.052) SD=6.75 (bias=~-0.00, SE=0.035)	[-2.73, -2.53]			

The results of the MRCP(UK) Highfliers were also compared with the results of those who did not make it to the Highfliers group due to the fact that they needed more attempts at any of the parts. Three such groups were distinguished: those who needed more attempts at Part I, but passed Part II and PACES at their first attempt ( $n=2,948$ ), those who needed more attempts at Part II, but passed Part I and PACES at their first attempt ( $n=381$ ), and those who needed more attempts in PACES, but passed Part I and Part II at their first attempt ( $n=4,186$ ). The comparisons always fell in favour of the MRCP(UK) Highfliers, as they in each case scored higher, and the differences in scores were highly significant each time (with  $p<0.001$ ). This supports the use of the Highfliers group as a reference group of high ability in further analyses.

#### **4.5.2 Differences based on demographic characteristics**

Previous studies found differences between groups of candidates with respect to certain demographic characteristics, which indicated the need to explore if such differences could also be found in the MRCP(UK) data available for this research. The factors pointed out in the previously referenced studies (see Chapter 2, section 2.4.1) to have an effect on the MRCP(UK) scores were: sex and declared ethnicity. This list of significant demographics was extended by two more factors: graduation from a UK university, otherwise referred to as Primary Medical Qualification ('PMQ'), and the fact of being trained in the UK. The reason for this is that the above-mentioned studies were based on a sample of UK graduates only, while this research used scores of all candidates who attempted MRCP(UK) between May 2003 and January 2011 including IMGs (international graduates).

##### **4.5.2.1 Sex**

Previous studies have found a difference between the performance of male and female candidates in the MRCP(UK) parts. Women outperformed men in PACES (Dewhurst *et al.*, 2007), while men obtained higher scores than women in Part I (McManus, Elder, *et al.*, 2008).

The parametric independent samples t-tests, bootstrapped independent samples t-test, and a non-parametric Mann-Whitney U test were performed on Part I, Part II and PACES scores in the History File with candidate's sex being the factor. Table 17 (next page) summarizes the means, results of the Mann-Whitney U, results of the t-tests (bootstrapped), and the effect sizes from non-parametric tests.

**Table 17. Comparison of mean scores at MRCP(UK) parts between groups based on sex (independent samples t-test results, also bootstrapped, and Mann-Whitney Z results) with effect sizes.**

<i>Group</i>	<i>N</i>	<i>Means and SDs</i>	<i>Bootstrapped Means and SDs</i>	<i>95% CI for the mean</i>	<i>Mann-Whitney Z</i>	<i>Independent samples t-test</i>	<i>Effect size (r)</i>
<b><u>Part I</u></b>							
Male	20,135	-4.40 SD=12.45	-4.40 (bias=-0.002, SE=0.090) SD=12.45 (bias=~0.00, SE=0.057)	[-4.58, -4.23]	Z=-5.96, p<0.001	t(35,954)=-6.62 p<0.001	0.04
Female	15,821	-3.56 SD=11.33	-3.56 (bias=0.001, SE=0.093) SD=11.33 (bias=0.001, SE=0.060)	[-3.73, -3.37]			
<b><u>Part II</u></b>							
Male	13,134	2.86 SD=7.95	2.86 (bias=0.001, SE=0.070) SD=7.95 (bias=~0.00, SE=0.056)	[2.73, 2.99]	Z=-10.24, p<0.001	t(22,396)=-9.55, p<0.001	0.07
Female	9,264	3.85 SD=7.08	3.85 (bias=0.001, SE=0.074) SD=7.08 (bias=~0.00, SE=0.057)	[3.69, 3.99]			
<b><u>PACES</u></b>							
Male	12,474	-2.67 SD=7.09	-2.67 (bias=0.001, SE=0.065) SD=7.09 (bias=-0.001, SE=0.041)	[-2.81, -2.54]	Z=-36.83, p<0.001	t(21,258)=-38.13 p<0.001	0.25
Female	8,786	0.89 SD=6.11	0.89 (bias=0.001, SE=0.066) SD=6.11 (bias=0.001, SE=0.049)	[0.75, 1.01]			

Inspection of Table 17 shows that women performed significantly better than men; this effect was statistically significant irrespective of the method employed. However, the effects should be considered weak to moderate.

#### **4.5.2.2 Ethnicity**

Previous studies have found that doctors from ethnic minorities underperformed in several measures of academic performance in comparison to their white colleagues (McManus, Richards, Winder, & Sproston, 1998; Woolf, Potts, & McManus, 2011; Yates & James, 2006). This was also found for MRCP(UK) (Dewhurst *et al.*, 2007). However, the referenced studies looked at the UK graduates only. The potential differences present in the History File were studied using parametric independent samples t-test, also bootstrapped, and univariate non-parametric comparison of means (Mann-Whitney Z). The results are summarised in Table 18 (next page).

The results indicated that white candidates scored significantly higher than their non-white colleagues in all parts of MRCP(UK), which is in concordance with the results of previous studies. Bootstrapped parametric independent samples t-test corroborated the results. The observed effects should be classified as moderate based on the value of the calculated effect sizes.

**Table 18. Comparison of mean scores in MRCP(UK) parts between groups based on declared ethnicity (independent samples t-tests, also bootstrapped, and Mann-Whitney Z results) with effect sizes.**

<i>Group</i>	<i>N</i>	<i>Means and SDs</i>	<i>Bootstrapped Means and SDs</i>	<i>95% CI for the mean</i>	<i>Mann-Whitney Z</i>	<i>Independent samples t-test</i>	<i>Effect size (r)</i>
<b><u>Part I</u></b>							
Non-white	25,785	-5.73 SD=12.16	-5.73 (bias=0.00, SE=0.076) SD=12.16 (bias=0.00, SE=0.048)	[-5.88, -5.58]	Z=-42.57, p<0.001	t(35,960)=- 43.93, p<0.001	0.25
White	10,177	0.27 SD=10.30	0.27 (bias=0.00, SE=0.100) SD=10.30 (bias=0.00, SE=0.073)	[0.07, 0.47]			
<b><u>Part II</u></b>							
Non-white	15,266	1.80 SD=7.53	1.80 (bias =0.001, SE=0.063) SD= 7.52 (bias =0.00, SE=0.048)	[1.68, 1.92]	Z=-42.84, p<0.001	t(22,396)=-44.07, p<0.001	0.31
White	7,132	6.41 SD=6.81	6.41 (bias =0.001, SE=0.080) SD= 6.81 (bias =0.001, SE=0.065)	[6.26, 6.58]			
<b><u>PACES</u></b>							
Non-white	14,324	-2.96 SD=6.92	-2.96 (bias =0.004, SE=0.055) SD= 6.92 (bias =0.001, SE=0.039)	[-3.07, -2.84]	Z=-54.42, p<0.001	t(21,258)=- 57.46, p<0.001	0.40
White	6,936	2.45 SD=5.34	2.45 (bias =0.002, SE=0.064) SD= 5.34 (bias =0.002, SE=0.049)	[2.32, 2.59]			

#### **4.5.2.3 Graduation from a UK university**

PMQ is a known factor affecting the performance of doctors. It was found that non-UK graduates tend to perform less well in training, for example, they receive lower scores in WBAs (Levy, Mohanaruban, & Smith, 2011b), and are more likely to receive a less satisfactory outcome in an ARCP assessment (Tiffin, Illing, Kasim, & McLachlan, 2014). International graduates were also found to be less likely to pass MRCP(UK) and MRCGP examinations (McManus & Wakeford, 2014), or pass PACES after a revision course (Bessant *et al.*, 2006). The differences between UK and international graduates ('IMGs') are often explained by previous training experience, cultural and ethical factors, and language proficiency (Esmail & Roberts, 2013; Slowther, Lewando Hundt, Taylor, & Purkis, 2009). The above-referenced literature, but also others (Hawtin, Williams, McKnight, & Booth, 2014), indicate that in some cases, there is also an interaction effect between being non-white and being an international graduate. Therefore, it was verified if PMQ was a significant factor for candidate scores through univariate parametric (bootstrapped), and non-parametric tests. The results of analyses are summarised in Table 19 (next page).

The non-parametric statistical tests showed that UK graduates scored higher than their colleagues who graduated abroad, and this effect was corroborated by the results of the parametric tests ( $p < 0.001$ ), and through bootstrapping ( $p < 0.001$ ). The observed effects should be considered moderate.

As described in the literature and based on the data from the History File, PMQ was found to be related to ethnicity. There was a significant correlation between being an IMG and being of non-white origins ( $\chi^2 (1, (n = 50,311)) = 16,992.56, p < 0.001, \Phi = 0.58, p < 0.001$ ). The odds of being non-white in the group of UK graduates equalled 0.71, while the odds of being non-white in the group of non-UK graduates were 15.8, which resulted in an odds ratio of 22.2. This meant that being non-white was 22.2 times more likely in the group of non-UK graduates.

The analyses of the relationships between PMQ, ethnicity, other demographic characteristics, and MRCP(UK) scores are presented in section 4.5.2.5, where the results of bootstrapped factorial ANOVA are provided.



**Table 19. Comparison of mean scores at MRCP(UK) parts between groups based on PMQ (independent samples t-tests, also bootstrapped, and Mann-Whitney Z results) with effect sizes.**

<i>Group</i>	<i>N</i>	<i>Means and SDs</i>	<i>Bootstrapped Means and SDs</i>	<i>95% CI for mean</i>	<i>Mann-Whitney Z</i>	<i>Independent samples t-test</i>	<i>Effect size (r)</i>
<b><u>Part I</u></b>							
Non-UK	21,741	-6.86 SD=12.26	-6.86 (bias=-0.001, SE=0.084) SD=12.26 (bias=0.002, SE=0.054)	[-7.03, -6.70]	Z=-54.51, p<0.001	t(35,960)=-57.96, p<0.001	0.30
UK	14,221	0.29 SD=10.09	0.29 (bias=-0.001, SE=0.083) SD=10.09 (bias=0.001, SE=0.058)	[0.12, 0.46]			
<b><u>Part II</u></b>							
Non-UK	12,025	1.06 SD=7.55	1.06 (bias=0.003, SE=0.067) SD=7.55 (bias=-0.002, SE=0.055)	[0.92, 1.19]	Z=-47.14, p<0.001	t(22,396)=-49.17, p<0.001	0.31
UK	10,373	5.83 SD=6.85	5.83 (bias=~0.00, SE=0.068) SD=6.85 (bias=-0.002, SE=0.053)	[5.69, 5.95]			
<b><u>PACES</u></b>							
Non-UK	10,384	-4.55 SD=6.73	-4.55 (bias=~0.00, SE=0.064) SD=6.73 (bias=~0.00, SE=0.045)	[-4.68, -4.42]	Z=69.17, p<0.001	t(21,258)=-78.20, p<0.001	0.47
UK	10,876	2.00 SD=5.44	2.00 (bias=~0.00, SE=0.052) SD=5.44 (bias=0.002, SE=0.036)	[1.90, 2.11]			

#### **4.5.2.4 Being a Probable UK trainee**

Being a probable UK trainee was a binary variable that aimed to capture a situation when someone might not have graduated from a UK university and might have been registered with the GMC and lived in the UK in order to get into a training position, e.g. through passing MRCP(UK). This variable was not based on relationships previously established in the literature, but rather was created to distinguish doctors who attempt MRCP(UK) with no intent of practicing in the UK from those who do intend to work in the UK.

Based on the data provided, it was found that being a non-UK trainee in the group of UK graduates was highly unlikely (odds=0.04), while being a non-UK trainee in the group of non-UK graduates was highly probable (odds=2.09). The odds ratio equalled 51.05, meaning that being a UK trainee was fifty times more likely in the group of UK graduates than in the non-UK graduates group. This effect was highly statistically significant with  $\chi^2(1, n=50,311)=19,727.34, p<0.001$  and  $\Phi = 0.63, p<0.001$ .

Independent samples t-test, bootstrapped t-test, and univariate non-parametric tests were performed to examine if being a probable UK trainee affected MRCP(UK) scores. The results of the analyses are summarised in Table 20 (next page). The results of the non-parametric tests showed clearly that doctors who were UK trainees scored significantly higher than their colleagues who trained elsewhere, and this effect was supported by parametric tests ( $p<0.001$ ) and bootstrapped parametric tests. The observed effects should be considered small to moderate in magnitude.

**Table 20. Comparison of mean scores at MRCP(UK) parts between groups based on being a probable UK trainee (independent samples t-tests, also bootstrapped, and Mann-Whitney Z results) with effect sizes.**

<i>Group</i>	<i>N</i>	<i>Means and SDs</i>	<i>Bootstrapped Means and SDs</i>	<i>95% CI for the mean</i>	<i>Mann-Whitney Z</i>	<i>Independent samples t-test</i>	<i>Effect size (r)</i>
<b><u>Part I</u></b>							
Non-UK	15,590	-5.84 SD=12.71	-5.84 (bias=0.00, SE=0.101) SD=12.71 (bias=0.001, SE=0.064)	[-6.04, -5.63]	Z=-22.70, p<0.001	t(35,960)=-25.32, p<0.001	0.13
UK	20,372	-2.64 SD=11.19	-2.64 (bias=-0.001, SE=0.080) SD=11.19 (bias=0.00, SE=0.051)	[-2.81, -2.49]			
<b><u>Part II</u></b>							
Non-UK	7,656	1.78 SD=7.91	1.78 (bias=0.002, SE=0.092) SD=7.91(bias=-0.003, SE=0.075)	[1.60, 1.96]	Z=-20.70, p<0.001	t(22,369)=-21.28, p<0.001	0.14
UK	14,742	4.04 SD=7.34	4.04 (bias=0.001, SE=0.064) SD=7.34 (bias=0.003, SE=0.046)	[3.91, 4.17]			
<b><u>PACES</u></b>							
Non-UK	5,869	-4.11 SD=7.13	-4.11 (bias=0.00, SE=0.094) SD=7.13 (bias=-0.001, SE=0.061)	[-4.28, -3.93]	Z=-36.37, p<0.001	t(21,258)=-39.19, p<0.001	0.26
UK	15,391	-0.09 SD=6.52	-0.09 (bias=-0.002, SE=0.051) SD=6.52 (bias=0.00, SE=0.036)	[-0.19, 0.02]			

#### 4.5.2.5 Combined effect of the demographic factors on MRCP(UK) scores

The above sections each presented a singular effect of a demographic factor on MRCP(UK) parts. The joint effect of all four factors was established with normal and bootstrapped factorial 2x2x2x2 ANOVA. The results from both methods were exactly alike. Non-parametric methods were omitted as they would have involved testing for sixteen ( $2^4$ ) pair-wise comparisons, and were unlikely to provide a clear analytical solution. Mean scores (with SDs) for individual groups were provided in the above sections, and therefore in order to avoid repetitiveness, they are not included in the tables of this section; the means for interaction terms are provided in Figures 12 to 16.

The summary of the analyses on the effect of demographic factors on Part I first attempt scores is presented in Table 21 (below).

**Table 21. Significant effects in a 2x2x2x2 factorial bootstrapped ANOVA on MRCP(UK) Part I scores with ethnicity, sex, PMQ and being a probable UK trainee as factors.**

<i>Factor</i>	<i>F-test value with significance</i>
Ethnicity	$F(1, 35,940) = 70.19, p < 0.001$
Sex	$F(1, 35,940) = 16.80, p < 0.001$
PMQ	$F(1, 35,940) = 371.95, p < 0.001$
Probable UK trainee	$F(1, 35,940) = 4.34, p = 0.037$
Sex * Ethnicity	$F(1, 35,940) = 5.02, p = 0.025$
Ethnicity * PMQ	$F(1, 35,940) = 4.54, p = 0.033$
PMQ * Probable UK trainee	$F(1, 35,940) = 48.96, p < 0.001$

Non-significant interaction terms were omitted in the table.

Inspection of Table 21 shows that all factors had a significant effect on Part I scores, and that a significant effect of certain interactions was present. Main effects were generally consistent with the results of analyses on individual factors; however, after taking into account other factors, female candidates scored significantly lower than male candidates (mean difference=-1.20 95%CI [-0.64, -1.72]; bias=-0.010, SE=0.279,  $p < 0.001$ ). White candidates scored higher than non-white candidates (mean difference=2.46 95%CI [1.90, 3.00], bias =0.014, SE=0.278,  $p < 0.001$ ), while UK graduates scored higher than IMGs (mean difference=5.66 95%CI [5.06, 6.21], bias=0.018, SE=0.289,  $p < 0.001$ ). Probable UK trainees obtained better results than non-UK trainees (mean difference=0.61 95%CI [0.06, 1.11], bias=0.016, SE=0.272,  $p = 0.022$ ), however, this effect was weaker than for the other three main factors.

The results of the analogous analysis with respect to Part II first attempt scores are presented in Table 22.

**Table 22. Significant effect in a 2x2x2x2 factorial bootstrapped ANOVA on MRCP(UK) Part II scores with ethnicity, sex, PMQ and being a probable UK trainee as factors.**

<i>Factor</i>	<i>F-test value with significance</i>
Ethnicity	$F(1, 22,382) = 155.49, p < 0.001$
Sex	$F(1, 22,382) = 11.45, p = 0.001$
PMQ	$F(1, 22,382) = 126.90, p < 0.001$
Sex*Ethnicity	$F(1, 22,382) = 5.87, p = 0.015$
PMQ * Probable UK trainee	$F(1, 22,382) = 18.55, p = 0.001$
Ethnicity * PMQ* Probable UK trainee	$F(1, 22,382) = 5.71, p = 0.017$

Non-significant interaction terms were omitted.

These results indicated that the only main effect that was statistically non-significant ( $F(1,22,382) = 0.71, p = 0.400$ ) was that of being a probable UK trainee; however, this factor influenced Part II scores indirectly, via interactions with PMQ, and with ethnicity and PMQ. Female candidates scored lower than male candidates after taking into account other factors (mean difference = -0.86 95%CI [-1.35, -0.34], bias = 0.005,  $SE = 0.262, p = 0.002$ ), white candidates scored higher than non-white candidates (mean difference = 3.16 95%CI [2.66, 3.69], bias = 0.003,  $SE = 0.262, p = 0.007$ ), and UK graduates scored higher than IMGs (mean difference = 2.85 95%CI [2.36, 3.35], bias = 0.005,  $SE = 0.260, p = 0.001$ ).

Analogous analyses performed for PACES scores showed that again the three main effects were significant, as were two of the interaction terms. The results are summarised in Table 23 (below).

**Table 23. Significant effects in a 2x2x2x2 factorial bootstrapped ANOVA on MRCP(UK) PACES scores with ethnicity, sex, PMQ and being a probable UK trainee as factors.**

<i>Factor</i>	<i>F-test value with significance</i>
Ethnicity	$F(1, 21,244) = 131.28, p < 0.001$
Sex	$F(1, 21,244) = 80.73, p < 0.001$
PMQ	$F(1, 21,244) = 341.74, p < 0.001$
Sex* Probable UK trainee	$F(1, 21,244) = 12.48, p < 0.001$
PMQ* Probable UK trainee	$F(1, 21,244) = 10.04, p = 0.002$

Non-significant interaction terms were omitted.

Female candidates scored significantly better in PACES than male candidates (mean difference = 2.00 95%CI [1.56, 2.43], bias = 0.007,  $SE = 0.222, p < 0.001$ ), white candidates

scored significantly higher than non-white candidates (mean difference=2.55 95%CI [2.15, 2.98], bias =0.003, SE=0.210,  $p=0.001$ ), and UK graduates scored significantly higher than IMGs (mean difference=4.12 95%CI [3.68, 4.56], bias=0.006, SE=0.217,  $p=0.001$ ).

The significant interactions between demographic factors included in the analyses were plotted as means to visualise the effects and establish which groups scored higher and which lower in the MRCP(UK) examinations.

An interaction between sex and ethnicity that was observed for Part I scores and Part II scores is presented as group means (with standard error of means) in Figure 12 (below).

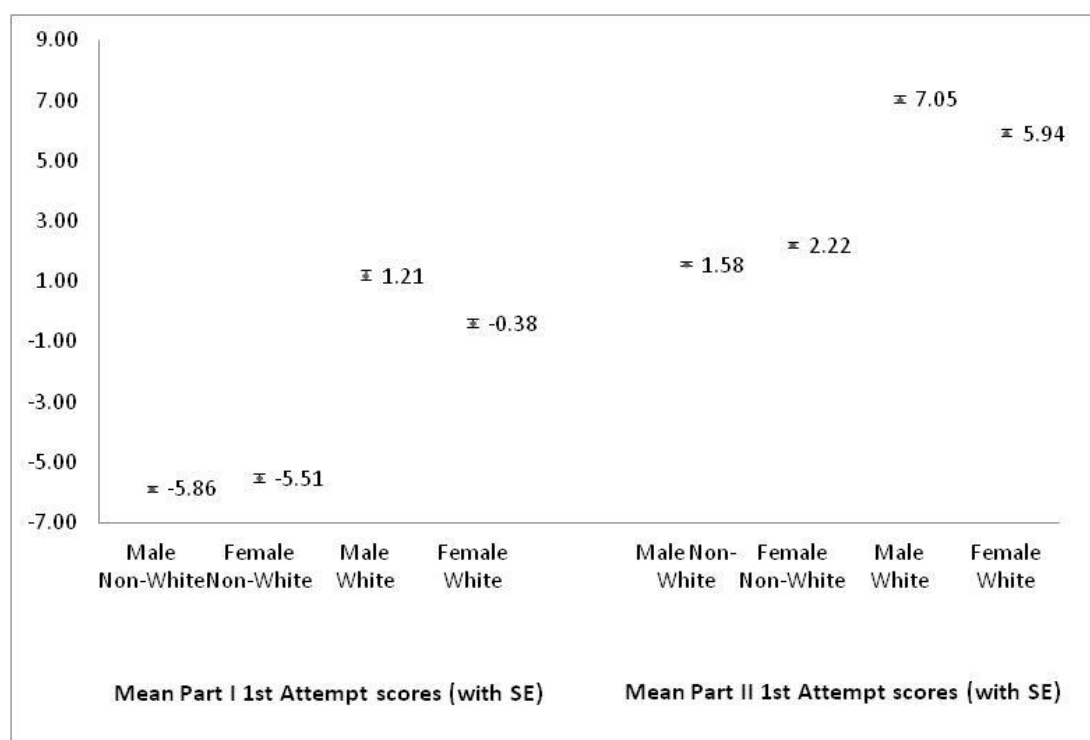


Figure 12. Mean scores (with SE) in Part I and Part II for groups based on sex and ethnicity (interaction effect).

White male candidates were better than all other groups, while non-white male candidates obtained the lowest scores in both Part I and Part II first attempt scores.

The interaction between ethnicity and PMQ observed in Part I scores was analogously plotted in Figure 13, and it was observed that white UK graduates scored significantly better than all other candidates, with non-white IMGs scoring the lowest.

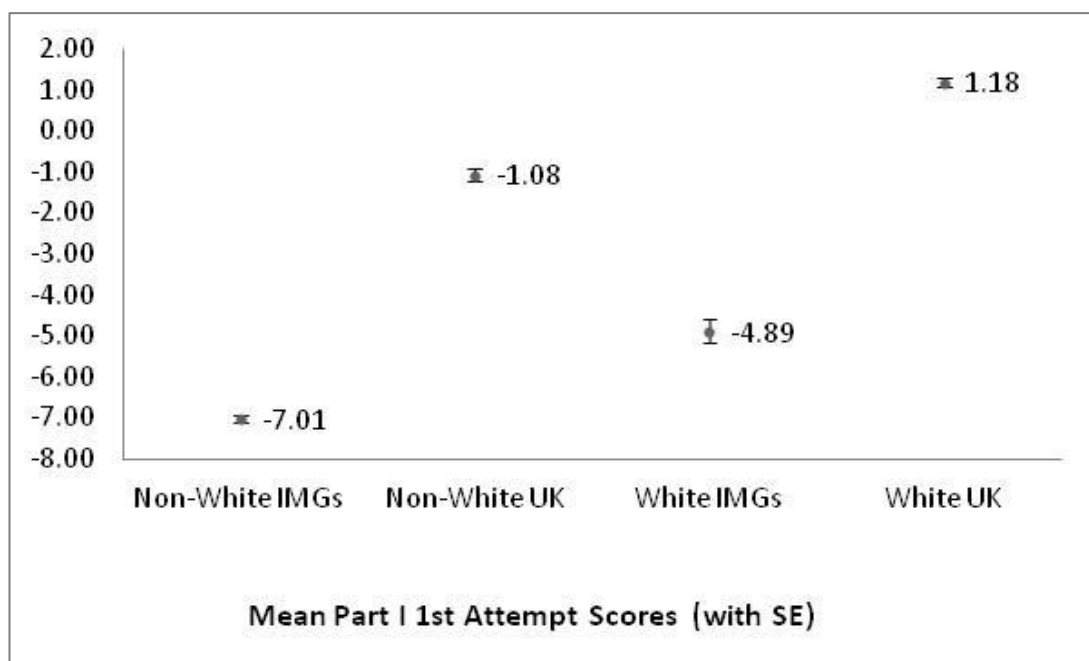


Figure 13. Mean scores (with SE) in Part I for groups based on PMQ and ethnicity (interaction effect).

Another significant observed interaction concerned the relationship between PMQ and being a probable UK trainee, which was present across all MRCP(UK) scores (see Figure 14, next page). The results clearly show that UK graduates score higher than IMGs, but the lowest scores are obtained by the UK trained IMGs.

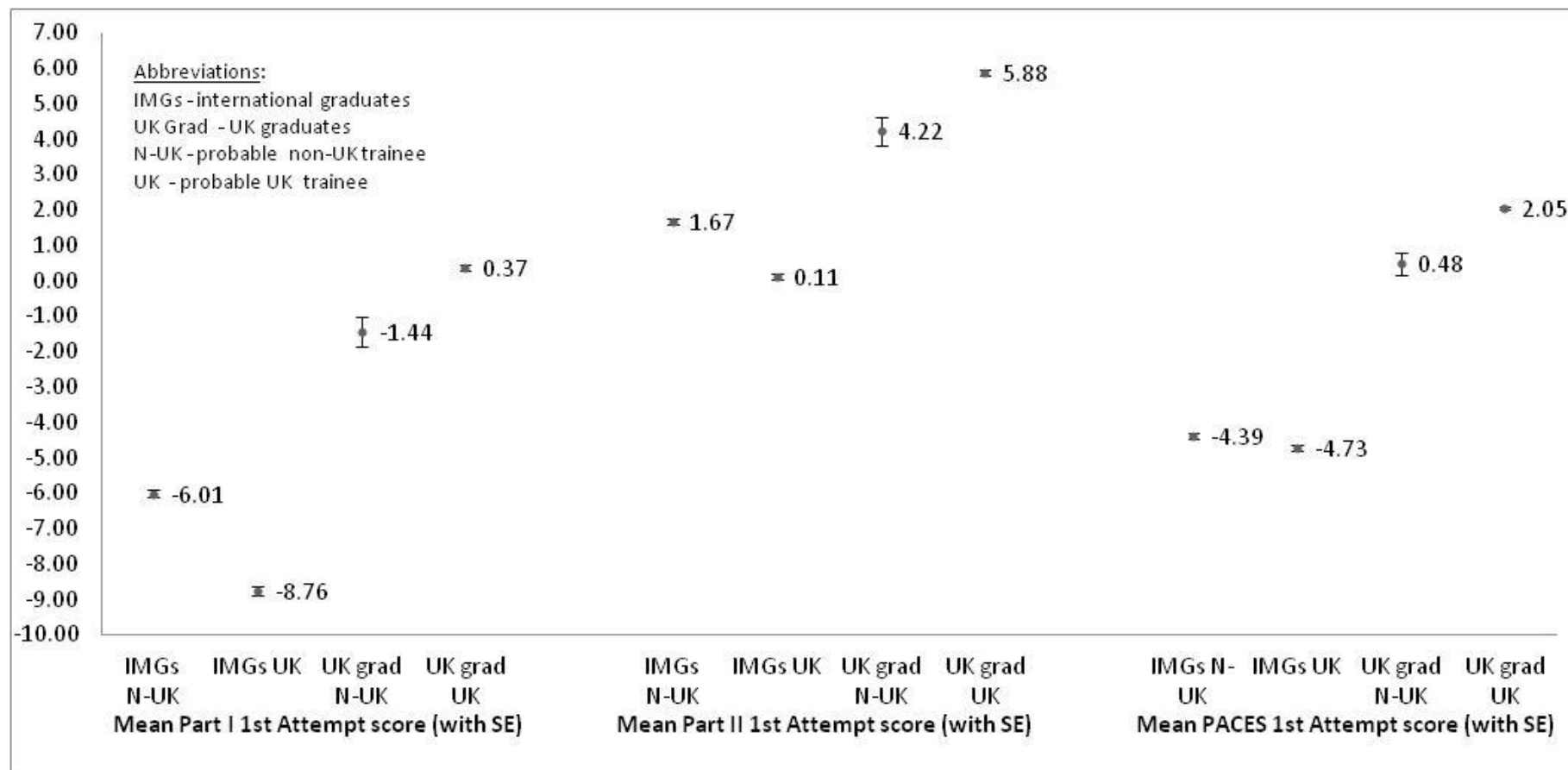


Figure 14. Mean scores (with SE) in all MRCP(UK) parts for groups based on PMQ and probable training (interaction effect).



The last interaction between two factors was observed for PACES scores, between sex and being a probable UK trainee. The means for the four groups are presented in Figure 15 (below).

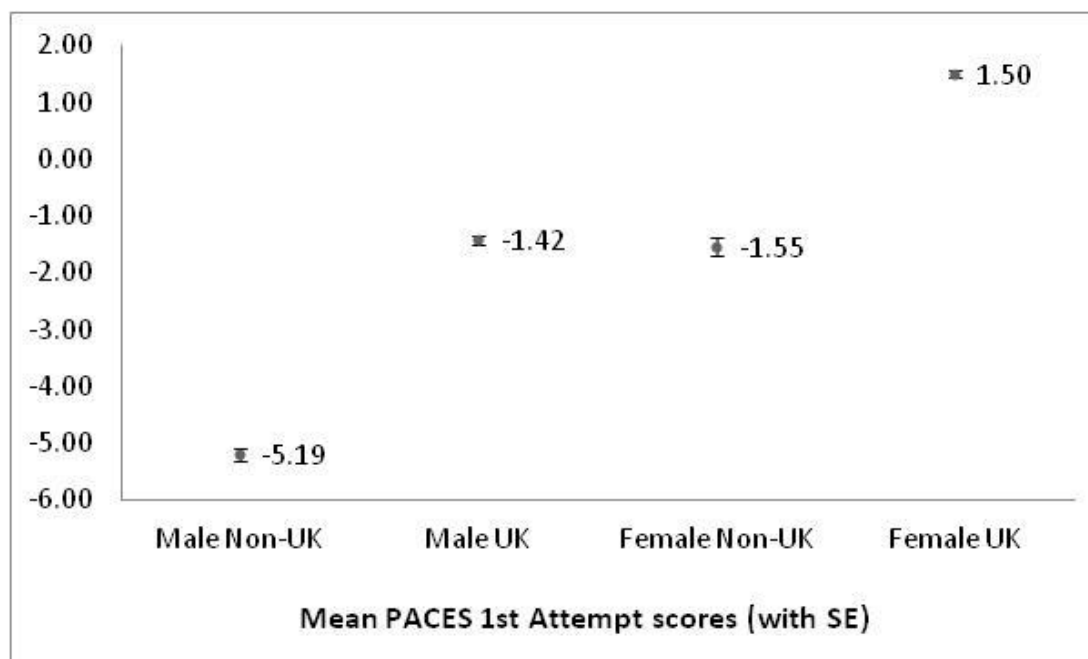


Figure 15. Mean scores (with SE) in PACES for groups based on sex and probable UK training (interaction effect).

The means indicate that female UK-trained doctors scored significantly higher than other groups, while female non-UK trained candidates and male UK trained doctors obtained almost equal scores. The lowest scoring group were the non-UK trained male doctors.

The final three-way interaction was observed for Part II first attempt scores between PMQ, ethnicity, and being a probable UK trainee. Three-way interactions are usually difficult to interpret; however, the means were sorted in an ascending order in Figure 16 (next page) to visualise the pattern.

Non-white candidates and IMGs scored generally lower than white UK graduates, irrespective of the training they received.

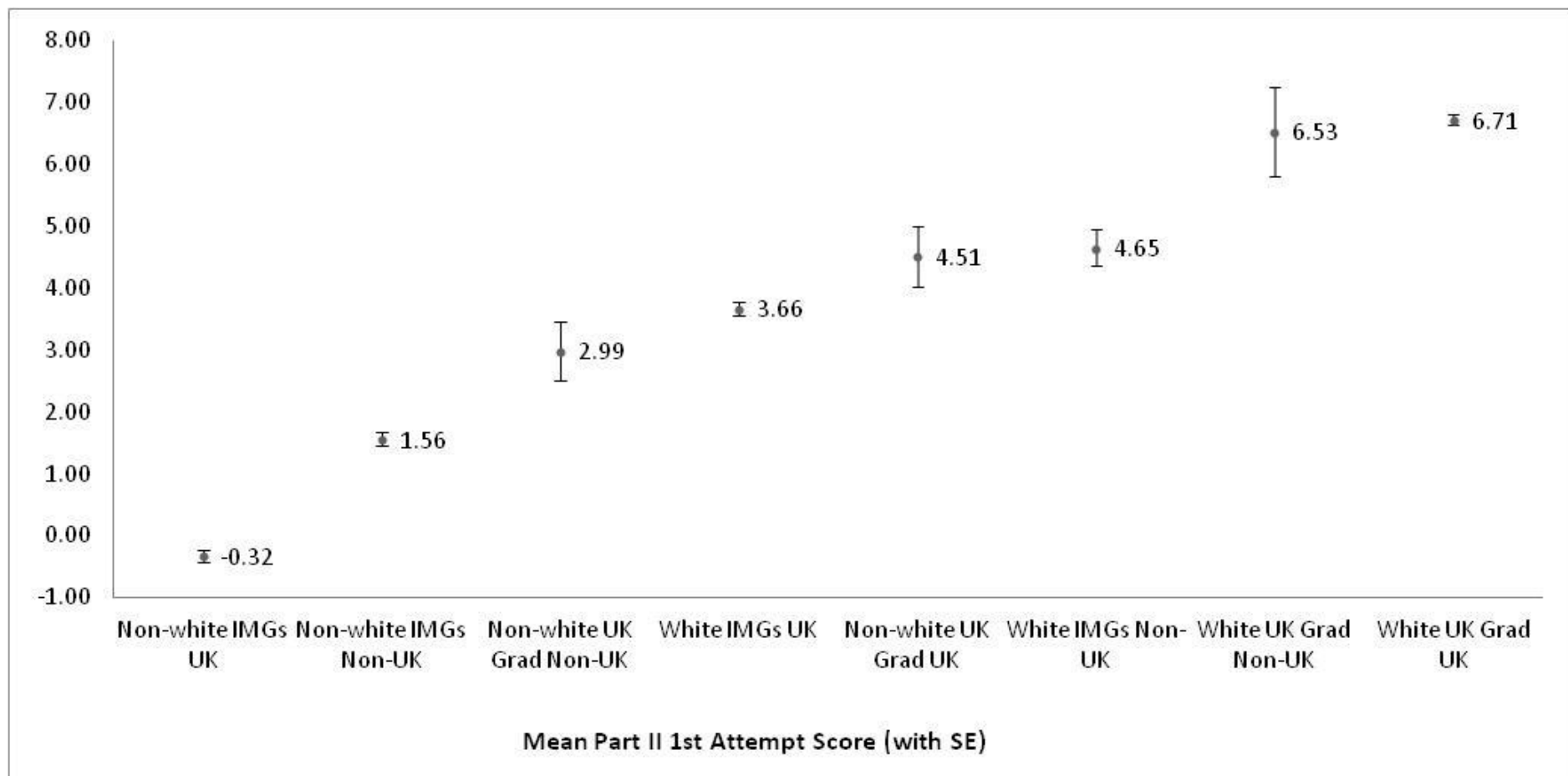


Figure 16. Mean scores (with SE) in Part II for groups based on ethnicity, PMQ and probable UK training (interaction effect).

## 4.6 REGRESSION MODELS

The previous analyses have shown that Part I, Part II, and PACES all show a relationship to sex, ethnicity, PMQ, and trainee status. The latter three are, however, correlated and therefore the question of whether the effects are all independent arises. The prediction models could have been estimated to find how the different parts predict one another. Further, based on the fact that the demographic characteristics of the candidates had an effect on their scores in MRCP(UK) parts, in order to estimate the joint effect of those characteristics, a linear regression model was fitted for each of the MRCP(UK) parts. Due to limitations resulting from violation of the assumption of normal distribution of the variables, the regressions were bootstrapped. Three models were fitted. The model explaining the Part I scores was based only on demographic predictors. The model explaining the Part II scores included the Part I scores and the demographic characteristics, while in the case of the model for PACES all: Part I scores, Part II scores, and the demographic predictors were included. This order in fitting the models was assumed from the order in which MRCP(UK) parts were taken until the end of 2008. The entry method was used.

Table 24 contains standardised and non-standardised bootstrapped coefficients for the regression models, number of valid cases based on which they were estimated, and explained variance.

The models indicated that ethnicity, sex, and being a UK graduate were significant predictors for all MRCP(UK) scores. Being a UK trainee was a significant predictor for Part I scores and an almost significant, however weak, predictor of Part II scores. Although sex was a significant factor, it had almost no impact on Part I and Part II scores based on the sizes of beta coefficients. However, in the case of PACES, the effect was present and should be considered weak, with the beta coefficient being equal to 0.13. The impact of ethnicity was constant for all models, varying between 0.10 and 0.12. Being a UK graduate had a significantly higher impact than ethnicity in the case of Part I and PACES, with beta coefficients being up to three times higher than those observed for ethnicity. There was a high and significant effect of Part I on Part II scores ( $\beta=0.57$ ), but the effects of Part I and Part II on PACES were much smaller ( $\beta=0.13$  and  $0.19$ ).

Further, the simple regression models and bootstrapped models yielded the exact same values of coefficients; therefore, they were not repeated in Table 24. The only observed differences between bootstrapped and simple models were the p-values associated with

the regression coefficients. Whenever a p-value was estimated as lower than 0.001 it became equal to 0.001 in the bootstrap procedure. In the case of non-significant effect of being a probable UK trainee, the p-values of the bootstrapped coefficients increased ( $p=0.052$  instead of  $p=0.051$  for Part II model, and  $p=0.872$  instead of  $p=0.851$  for PACES model).

Extending the above simple models, three hierarchical models were also fitted, where consecutive blocks were entered based on the size of the beta coefficients from the highest to the lowest, after taking into account the order in which MRCP(UK) parts are to be taken. The estimates of the coefficients remained the same.

**Table 24. Summary of the regression models fitted for Part I, Part II and PACES scores as dependent variables and demographic factors as predictors.**

Model	Const.	Ethnicity	Sex	Predictors		Part I score	Part II score
				UK grad	UK trainee		
<b><u>Dependent variable: Part I (n=35,956), R<sup>2</sup>=9.8%</u></b>							
Non-stand. coeff.	-5.04** bias=-0.009 SE=0.196 95% CI[-5.43,-4.66]	2.58** bias=-0.004 SE=0.155 95% CI [2.26, 2.88]	-0.91** bias=0.008 SE=0.122 95% CI[-1.14, -0.66]	7.61** bias=-0.006 SE=0.172 95% CI[7.26, 7.95]	-2.50** bias=0.003 SE=0.159 95% CI[-2.81, -2.18]	n/a	n/a
Stand. β coeff.s	n/a	0.10**	-0.04**	0.31**	-0.10**	n/a	n/a
<b><u>Dependent variable: Part II (n=16,744), R<sup>2</sup>=41.7%</u></b>							
Non-stand. coeff.	2.23** bias=-0.010 SE=0.151 95% CI[1.93, 2.51]	1.85** bias=~0.00 SE=0.110 95% CI[1.61,2.07]	-0.39** bias=0.004 SE=0.091 95% CI[-0.56,-0.21]	2.14** bias=~0.00 SE=0.127 95% CI[1.89, 2.39]	-0.24 <sup>(a)</sup> bias=0.005 SE=0.125 95% CI[-0.47, 0.02]	0.45** bias=~0.00 SE=0.005 95% CI[0.44, 0.46]	n/a
Stand. β coeff.s	n/a	0.12**	0.03**	0.14**	-0.02 <sup>(a)</sup>	0.57**	n/a
<b><u>Dependent variable: PACES (n=11,973), R<sup>2</sup>=31.1%</u></b>							
Non-stand. coeff.	-6.97** bias=0.002 SE=0.192 95% CI[-7.36, -6.61]	1.60** bias=0.003 SE=0.107 95% CI[1.33, 1.81]	1.78** bias=~0.00 SE=0.119 95% CI[1.57, 2.00]	3.88** bias=-0.004 SE=0.146 95% CI[3.59, 4.16]	0.03 bias=-0.002 SE=0.166 95% CI[-0.31, 0.34]	0.09** bias=~0.00 SE=0.007 95% CI[0.08, 0.11]	0.19** bias=~0.00 SE=0.011 95% CI[0.17, 0.21]
Stand. β coeff.s	n/a	0.11**	0.13**	0.28**	0.002	0.13**	0.19**

\*\*significant at  $p \leq 0.001$ , (a)  $p = 0.051$  – almost significant

In summary, the models seem to suggest that in the case of Part I being a UK graduate was more indicative of future exam results than ethnicity and sex. However, the negative beta coefficient for being a UK trainee is confusing. It is possible that the negative effect resulted from high correlations between being a UK trainee and ethnicity and PMQ. The models for Part II indicated that the best predictor were Part I results followed by ethnicity and UK PMQ. In the case of PACES, the best predictor was being a UK graduate, followed by Part II results, sex, ethnicity, and Part I results.

#### **4.7 STRUCTURAL EQUATION MODEL**

In order to estimate the effect of each of the predictors on Part I, Part II, and PACES scores jointly after taking into account the relationships between the parts themselves, a structural equation model was proposed. The coefficients for the model were estimated using IBM SPSS AMOS (IBM Corp., 2010). The model structure is presented in Figure 17.

The coefficients estimated for this model were very similar to the ones estimated for the simple regression models. Sex played a much bigger role for PACES scores, while for Part I and Part II scores its effect was almost zero. The effect of ethnicity on Part I, Part II, and PACES scores was consistent with coefficients equal from 0.09 to 0.010. PMQ was a significant predictor for all MRCP(UK) parts scores, with a varied impact from 0.11 on Part II scores to 0.30 on Part I scores. Being a probable UK trainee had almost no impact on MRCP(UK) scores with the exception of the Part I score, where the standardised coefficient reached -0.09, which might be confusing. The SEMo model, however, was not well-fitted based on the value of  $\chi^2(2)=2,776.62$ ,  $p<0.001$ , which was corroborated by the RMSEA value equal to 0.224 (higher than 0.08), which suggest caution in interpretation of the model's results. The standardised and non-standardised coefficients with *SE* and associated significance levels are summarised in Table 25.

The exogenous variables (sex, ethnicity, PMQ, and being a probable UK trainee) were intercorrelated. The estimates of the correlation coefficients are also presented in Table 25. They were all highly significant. There was a significant correlation between ethnicity and PMQ ( $r=0.58$ ), and PMQ and UK training ( $r=0.63$ , the estimate was similar to the one provided previously in section 4.5.2.4), as well as between ethnicity and UK training ( $r=0.43$ ). There were rather small correlations between being female and being a UK graduate and of white ethnicity.

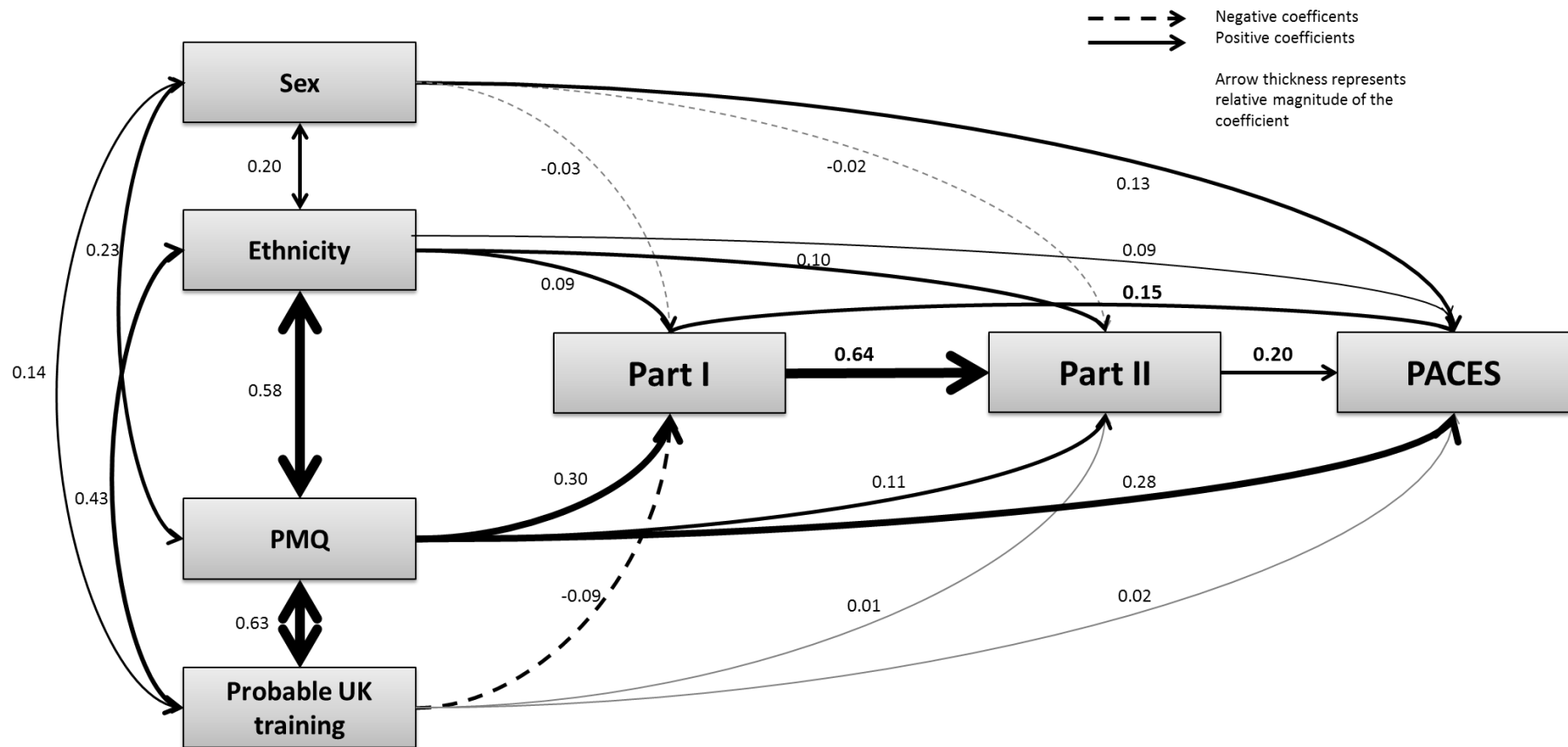


Figure 17. Structural Equation Model describing relationships between MRCP(UK) first attempt scores and demographic factors (sex, ethnicity, PMQ, and being a probable UK trainee);  $n = 50,311$ .

**Table 25. Standardised and non-standardised coefficients with significance levels for the structural equation model fitted for MRCP(UK) scores and demographic factors.**

<i>Relationship</i>			<i>Non- standardised coefficients</i>	<i>SE</i>	<i>p- value</i>	<i>Standardised Coefficients (<math>\beta</math>)</i>
Part1MarkAttempt1	←	PMQ (UK)	7.31	0.172	***	0.30
Part1MarkAttempt1	←	Probable UK trainee	-2.09	0.151	***	-0.09
Part1MarkAttempt1	←	Ethnicity (White)	2.46	0.164	***	0.09
Part1MarkAttempt1	←	Sex (Female)	-0.83	0.121	***	-0.03
Part2MarkAttempt1	←	PMQ (UK)	1.82	0.122	***	0.11
Part2MarkAttempt1	←	Ethnicity (White)	1.76	0.114	***	0.10
Part2MarkAttempt1	←	Probable UK trainee	0.17	0.105	0.116	0.01
Part2MarkAttempt1	←	Sex (Female)	-0.30	0.084	***	-0.02
Part2MarkAttempt1	←	Part1MarkAttempt1	0.43	0.004	***	0.64
PACESMarkAttempt1	←	Sex (Female)	1.88	0.083	***	0.13
PACESMarkAttempt1	←	Probable UK trainee	0.21	0.103	0.039	0.02
PACESMarkAttempt1	←	PMQ (UK)	3.99	0.122	***	0.28
PACESMarkAttempt1	←	Ethnicity (White)	1.46	0.113	***	0.09
PACESMarkAttempt1	←	Part1MarkAttempt1	0.09	0.006	***	0.15
PACESMarkAttempt1	←	Part2MarkAttempt1	0.18	0.008	***	0.20
<b><i>Correlations</i></b>						
Sex (Female)	↔	Ethnicity (White)			***	0.20
PMQ (UK)	↔	Ethnicity (White)			***	0.58
Probable UK trainee	↔	PMQ (UK)			***	0.63
Probable UK trainee	↔	Ethnicity (White)			***	0.43
Sex (Female)	↔	PMQ (UK)			***	0.23
Sex (Female)	↔	Probable UK trainee			***	0.14

\*\*\*  $p < 0.001$

## SUMMARY AND DISCUSSION

The analyses presented in this chapter aimed to test if MRCP(UK) parts' scores predict one another in accordance with the general hypotheses of this thesis. Further, this chapter aimed to present the existing relationships observed in the data of the History File. Among others, it described the relationships between the first attempt results and consecutive attempts at the same MRCP(UK) part, which led to employing first attempt results as the



key measures of performance. Further, it described the relationship between first attempt results in Part I, Part II, and PACES and the effects of demographic characteristics on MRCP(UK) parts results.

The findings regarding the relationship between first attempt scores in the three MRCP(UK) parts supported the general hypothesis. Part I scores predicted scores in Part II and PACES, although to a different degree, and Part II scores predicted PACES scores. The strongest relationship was observed between Part I and Part II ( $r=0.61$ , uncorrected), while correlations between Part I or Part II and PACES were smaller ( $r=0.30$  and  $r=0.38$ , respectively). These correlations support the notion of the predictive validity of MRCP(UK). At the same time, the strength of the coefficients would not suffice to justify the aggregation of the results of individual parts into one measure.

The strength of those uncorrected validity coefficients was in line with psychometric theory and the assumed hypotheses. As previously mentioned, raw validity coefficients rarely exceed 0.60 (Cronbach, 1970, p.135). This is due to the fact that uncorrected validity coefficients are influenced by various factors, such as unreliability of the measures, the constructs underlying the design of the exams, the forms of exams, and the time-span between the exams (Anastasi & Urbina, 1997; Cronbach, 1970).

The unreliability of the measures was taken into account during the process of correcting for range restriction which requires disattenuation, i.e. correcting for unreliability. The process increased the values of the obtained validity coefficients; however, it did not affect the pattern of the relationships. In particular it did not result in the perfect alignment of the measures (coefficients being equal to 1), which suggests that Part I, Part II, and PACES measure different constructs, which overlap only to a certain extent that is equal to the value of the squared coefficient (common variance). Indeed, Part I and Part II aim to test knowledge from the same domain, i.e. medical knowledge, while PACES aims to test clinical ability and attitudes. The difference between Part I and Part II lies in Part I being more factual, while Part II tests for data interpretation based on knowledge. This would explain a moderately high correlation between Part I and Part II, which was still higher than the correlation coefficients associated with PACES.

The observed pattern also seems to suggest that both the forms of the exam and the time interval between them might have influenced the strength of the relationships. As argued by Cronbach (Cronbach, 1970, p.137), the time interval between the test and the criterion measurement is a factor that needs to be taken into account during the validation process;

the longer the time interval, the smaller the correlation. The time interval between Part I and Part II was on average sixty-one weeks, while that between Part I and PACES was a hundred and one weeks, which should result in the correlation coefficient between Part I and PACES being lower than that between Part I and Part II. Indeed, this was the case (corrected  $r=0.43$  and  $r=0.78$ , respectively). In accordance with the theory, the effect of the form of the exams would yield similar results. Part I and Part II are written exams which suggest the relationship between them should be stronger than the relationship between both written parts and PACES, as a practical clinical skills assessment.

The observed pattern of associations most likely resulted from all of the above factors; however, the underlying key factor is the relationship between constructs, as neither of the other factors would be sufficient to yield coefficients of such strength as presented. Therefore, the results of the correlation analyses between MRCP(UK) parts were supported by the associated psychometric theory of validity.

Consecutive analyses on the influence of demographic characteristics on MRCP(UK) performance resulted in confirming the significant effects of sex, ethnicity, and place of primary medical qualification, and in the case of Part I, also of being a probable UK trainee. These results were in line with the previous research, which found similar dependencies in undergraduate and postgraduate training performance (Dewhurst *et al.*, 2007; Haq, Higham, Morris, & Dacre, 2005; McManus, Woolf, & Dacre, 2008; McManus & Richards, 1996; Woolf *et al.*, 2011). The results showed that based on the data from the History File women scored higher than men in PACES, but lower in the other two parts of MRCP(UK). A significant difference was also observed in favour of white candidates, UK graduates, and UK trainees in comparison to non-white candidates, non-UK graduates, and non-UK trainees, respectively. However, it was also observed that being of non-white origin was highly correlated with being educated and trained abroad. Further, the linear regression models seemed to suggest that being a UK graduate was more relevant for future results of the MRCP(UK) examination than ethnicity or place of training or sex based on the values of beta coefficients. This may lead to a conclusion that the MRCP(UK) examination is heavily embedded in the British medical education system, and its principles, methods of teaching, and methods of progress assessment may present an obstacle for an international doctor. Early immersion into a British educational setting may facilitate acquiring tacit knowledge, or exposure to the hidden curriculum as it is sometimes referred to (Hafferty & Franks, 1994; Lempp & Seale, 2004), which is useful for preparing and passing MRCP(UK). For example, the familiarity of the setting may result in the candidate being more relaxed and

more focused. Such familiarity may also include the form of the test or a proper structure of argumentation in discussion. The effect of language proficiency may arguably be a meaningful factor, which should facilitate the process of passing the exam both in clinical and written test settings. In fact, it has been previously found that non-white candidates tend to be assessed more leniently in PACES stations by non-white examiners, where the mark relies heavily on communications skills (Dewhurst *et al.*, 2007). Although observed, the underlying reasons for such differences are speculative and not yet entirely clear. In fact, the impact of demographic characteristics on, among others, performance in MRCP(UK) is a subject of another research (Unwin, n.d.).

The results of the presented analyses were consistent irrespective of the analytical approach employed; the non-parametric tests, parametric tests, and bootstrapped methods yielded analogous results leading to analogous interpretations. Non-parametric tests would be a methodologically pure approach in all cases where the assumptions for the use of parametric tests were not held. However, they are considered less robust and limited in terms of analytical solutions available to the researcher (Gaito, 1959). The alternative approach was to ignore certain flaws in the data, provide parametric solutions, and treat them with caution; or to employ distribution-free bootstrapping to obtain the effect sizes and their significance. The comparison between these three methods suggested that despite the violation of certain assumptions, such as normality of the distribution, the parametric methods lead to identical conclusions about the nature of the relationships between the variables as the non-parametric ones. Therefore, subsequent analyses on relationships between MRCP(UK) and the chosen criteria employed parametric methods, because they were proven in this chapter to be robust and they offer the largest variety of tools for analysing multivariate relationships.

To summarize, the obtained results had a significant impact on subsequent analyses. The evidence presented in Chapter 4 allowed for use of first attempt scores as the key measure of performance due to the fact that they are predictive of all consecutive attempts. This effectively simplified the inferential process, as the analyses of all other attempts scores could be omitted. Further, the analyses presented above led to a conclusion that despite the observed associations between MRCP(UK) parts they should be treated as separate predictors, and therefore all analyses presented in Chapters 5 to 7 follow this approach.

## Chapter 5

### Relationship between MRCP(UK) and Measures of Knowledge

#### ABSTRACT

*It was hypothesised that MRCP(UK) would be predictive of performance in subsequent knowledge tests, which would provide evidence that it indeed measures medical knowledge. This would constitute evidence for its predictive validity. The sources of criterion data that were identified for that purpose were as follows: the Specialty Certificate Exams ('SCEs'), Cardiology Knowledge Based Assessment ('CKBA'), Clinical Oncology First and Final Exam ('FRCR1' and 'FRCR2') and the General Practitioners' Applied Knowledge Test ('AKT'), altogether comprising sixteen exams. The relationships between MRCP(UK) scores and measures of performance in those exams were investigated and the findings supported the hypothesis. It was found that MRCP(UK) part scores correlated with scores in the above-mentioned exams. Linear regression models showed Part II to be the best predictor in the majority of cases, and Part I to be the second best predictor. The effect of PACES on knowledge exams scores was not straightforward to interpret; however, whenever significant, the coefficients associated with PACES were smaller in magnitude when compared to the other two parts. Despite observed differences between the linear regression models, they were statistically similar as tested using multilevel modelling and Chow tests, with the only exception of AKT. A summary and discussion of the results is provided.*

The hypothesis for this research was that MRCP(UK) being a valid exam would predict performance in subsequent knowledge exams and clinical skills assessments. Several sources of comparison data were secured, as described in Chapter 2. This chapter focuses only on establishing the relationships between MRCP(UK) parts scores and scores in subsequent knowledge exams. The criterion measures therefore involved the scores obtained by MRCP(UK) candidates in the twelve specialty certificate exams ('SCEs'), the Cardiology specialty exam ('CKBA'), the Clinical Oncology specialty First and Final exams ('FRCR1' and 'FRCR2', excluding the clinical component of the FRCR2) and the MRCGP AKT exam. The SCEs, FRCR and CKBA examinations were all specialty examinations, meaning that they were attempted by the MRCP(UK) candidates several years (from four to six years) after completing MRCP(UK). The MRCGP AKT exam could have been attempted by

the MRCP(UK) candidates almost concurrently; however, all scores obtained pre-MRCP(UK) were excluded from the analyses. All of the above-mentioned exams were considered suitable sources of robust data, as described in section 2 of Chapter 3, based on the reliability coefficients and the validity evidence published in the professional literature. The effective sample sizes varied between the datasets, with Palliative Medicine providing the smallest sample ( $n=31$ ) for analysis, and MRCGP AKT providing the largest sample ( $n=7,685$ ). The detailed demographic composition of these datasets and the inferential analyses testing the hypothesis are presented in the following sections of this chapter.

## **5.1 SPECIALTY CERTIFICATE EXAMS**

Specialty Certificate Exams ('SCEs') is the joint name for final exams in twelve specialties administered by the RCP, namely: Acute Medicine, Dermatology, Endocrinology, Gastroenterology, Geriatric Medicine, Infectious Diseases, Neurology, Medical Oncology, Renal Medicine, Respiratory Medicine, Rheumatology, and Palliative Medicine. They are usually taken in the penultimate year of the higher Specialty Training, which is four to six years after MRCP(UK). These are written exams and test for applied medical knowledge and data interpretation, and as such, they resemble Part II. They consist of two papers with two hundred questions in total. In that sense they are similar to Part I examination, as Part II is a longer exam. More details can be found in section 2.4.4. Due to the fact that SCE exams are annually taken by a small number of candidates and that they are relatively new exams – they started in 2008 – the results within each specialty were analysed jointly without referring to particular years and reliability coefficients of these exams were averaged within each specialty across years.

### **5.1.1 Descriptive statistics**

Altogether, the RCPs provided information on 2,244 individual doctors who attempted at least one SCE since 2008. The dataset merged with the History File contained only 2,224 records which provided 2,542 valid scores. The majority of the candidates passed the chosen SCE. The average pass rate was 71.6% and varied across specialties. The majority of candidates also passed an SCE on their first attempt (1,443 candidates; 65.3%). Table 26 presents detailed distributions of recorded attempts and an overall pass-rate for each specialty separately.

**Table 26. Records of attempts and overall pass-rates by specialty in Specialty Certificate Exams.**

<i>Specialty</i>	<i>Number of attempts</i>				<i>Overall Pass-rate</i>
	<i>1st</i>	<i>2nd</i>	<i>3rd</i>	<i>in total</i>	
Acute Medicine	215	26	0	241	68.1%
Dermatology	164	11	1	176	88.8%
Endocrinology	295	56	1	352	64.5%
Gastroenterology	345	65	0	410	69.3%
Geriatric Medicine	304	26	0	330	77.5%
Infectious Diseases	48	4	1	53	70.9%
Neurology	149	17	1	167	75.0%
Medical Oncology	104	19	0	123	66.9%
Renal Medicine	168	36	2	206	64.0%
Respiratory Medicine	246	41	2	289	66.9%
Rheumatology	148	9	0	157	81.2%
Palliative Medicine	38	0	0	38	78.9%
<b>Total:</b>	<b>2,224</b>	<b>310</b>	<b>8</b>	<b>2,542</b>	<b>71.6%</b>

Demographic characteristics were available for 2,083 candidates. Among those, 1,267 were male (816 female), forming 61% of the sample. The majority of the candidates were non-white (64.2%). Half of the candidates in the dataset qualified at a British university (1,048; 50.3%), and most of them were UK trainees (1,758; 84.4%). The demographics varied depending on the specialty and Table 27 provides an overview for each specialty separately (as frequency of valid cases).

**Table 27. Demographic characteristics of the SCE candidates within each specialty (as per cent of total number of cases).**

<i>Specialty (n valid cases)</i>	<i>Sex (Male)</i>	<i>Ethnicity (Non-white)</i>	<i>UK PMQ</i>	<i>UK trainee</i>
Acute Medicine (n=209)	72.2%	65.1%	45.9%	85.2%
Dermatology (n=151)	38.4%	58.3%	68.9%	91.4%
Endocrinology (n=273)	64.5%	85.7%	18.7%	67.8%
Gastroenterology (n=333)	73.0%	67.0%	51.1%	83.8%
Geriatric Medicine (n=296)	53.7%	50.0%	65.5%	97.3%
Infectious Diseases (n=44)	43.2%	50.0%	65.9%	79.5%
Neurology (n=127)	67.7%	59.8%	58.3%	83.5%
Medical Oncology (n=87)	56.3%	56.3%	52.9%	75.9%
Renal Medicine (n=155)	72.3%	72.9%	34.8%	69.7%
Respiratory Medicine (n=235)	60.0%	57.9%	60.0%	92.8%
Rheumatology (n=139)	48.9%	74.1%	44.6%	89.9%
Palliative Medicine (n=34)	14.7%	29.4%	79.4%	94.1%

The first attempt scores in each of the SCEs were normally distributed, as confirmed by the values and significance of the K-S tests (see Table 28). Table 28 also contains the descriptive statistics for each SCE separately. Within 2,224 records present in the dataset, it was found that 148 cases did not have an MRCP(UK) record between 2003 and 2011, which further limited the sample to 2,076 valid cases in inferential analyses. The MRCP(UK) scores were also tested for normality within the sample resulting from the merge. Within each specialty the scores were distributed normally, allowing the use of parametric statistics in further analyses.

**Table 28. Distribution parameters for each SCE examination scores with one- sample K-S test results.**

<i>Specialty</i>	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>Range</i> (Min to Max)		<i>Skewness</i>	<i>Kurtosis</i>	<i>K-S test result</i>
Acute Medicine	215	63.06	8.56	34.74	82.65	-0.75	0.94	$Z_{KS} = 1.22, p=0.099$
Dermatology	164	77.02	6.87	51.50	90.00	-0.76	0.88	$Z_{KS} = 0.95, p=0.323$
Endocrinology	295	64.60	9.10	39.00	88.50	-0.14	-0.23	$Z_{KS} = 1.03, p=0.240$
Gastroenterology	345	66.27	6.86	42.93	80.00	-0.30	-0.24	$Z_{KS} = 0.79, p=0.556$
Geriatric Medicine	304	63.85	6.03	44.16	78.00	-0.39	0.14	$Z_{KS} = 0.83, p=0.498$
Infectious Diseases	48	68.70	12.59	34.00	90.50	-0.65	0.74	$Z_{KS} = 0.71, p=0.701$
Neurology	149	60.79	9.64	31.82	82.32	-0.57	0.30	$Z_{KS} = 0.86, p=0.444$
Medical Oncology	104	58.16	7.97	33.33	79.80	-0.10	0.73	$Z_{KS} = 0.51, p=0.960$
Renal Medicine	168	65.80	7.72	44.00	82.50	-0.44	0.04	$Z_{KS} = 0.88, p=0.425$
Respiratory Medicine	246	62.40	8.05	39.20	79.00	-0.13	-0.40	$Z_{KS} = 0.72, p=0.677$
Rheumatology	148	79.05	7.93	59.28	95.48	-0.34	-0.60	$Z_{KS} = 0.98, p=0.289$
Palliative Medicine	38	70.46	6.75	52.55	82.14	-0.50	0.38	$Z_{KS} = 0.77, p=0.597$
<b>Total:</b>	2,224	--	--	--	--	--	--	--



Due to specialties differing in terms of: a) parameters of the distributions of the results, b) corresponding pass-marks being set with the Hofstee method (Norcini, 2003a) for each specialty separately, c) pass-rates, and d) due to the fact that there were no anchor questions common to all specialties, the scores within each specialty were not directly comparable and could not have been statistically equated. This lack of comparability between the SCE results made it impossible to establish the difficulty rank for the specialties. Therefore, Z-transformed MRCP(UK) scores were used to measure the standard of the SCEs candidates for each specialty by employing one-way ANOVA. The assumption of homogeneity of variance was tested, and it was found to be valid for Part I and PACES scores only. Nonetheless, the analyses with regards to Part II results are also discussed below<sup>9</sup>.

One-way ANOVA tests showed that there were significant differences in MRCP(UK) performance between candidates of different specialties ( $F(11, 1,449) = 2.90, p=0.001$  for Part I;  $F(11, 1,840) = 3.43, p<0.001$  for Part II; and  $F(11, 2,043)=6.23, p<0.001$  for PACES). Post hoc REGW Q tests identified homogenous groups based on the mean specialty scores. In the case of Part I, all specialties turned out similar ( $p=0.063$ ); however, there was a large span between the average lowest and highest mean results, contributing to a significant ANOVA test. The lowest mean was observed for Acute Medicine (-0.24 SD; -0.15 percentage points), and the highest was observed for Infectious Diseases (+0.46 SD; +5.70 percentage points). In the case of Part II, two homogenous groups were observed. The first group included almost all specialties without Neurology and Infectious Diseases ( $p=0.068$ ). The second homogenous group ( $p=0.240$ ) contained all specialties without Endocrinology (-0.24 SD below the mean). PACES divided specialties into three groups. The first homogenous group ( $p=0.569$ ) comprised Endocrinology (-0.31 SD), Renal Medicine, Acute Medicine, Rheumatology, and Medical Oncology (-0.05 SD). The second homogenous group ( $p=0.183$ ) comprised Renal Medicine, Acute Medicine, Rheumatology, Medical Oncology, Gastroenterology, Geriatric Medicine, Respiratory Medicine, and Neurology (+0.14 SD). The third homogenous group ( $p=0.122$ ) comprised Rheumatology (-0.09 SD), Medical Oncology, Gastroenterology, Geriatric Medicine, Respiratory Medicine, Neurology, Dermatology, Infectious Diseases, and Palliative Medicine (+0.49 SD). Figures 18 and 19 show mean results for each specialty in Part II and PACES with the composition of homogenous groups.

---

<sup>9</sup> Part II results of the Levene's test:  $W(11, 1840) = 1.852, p=0.041$ ; the non-parametric Kruskal-Wallis H test performed for sensitivity purposes was highly significant ( $\chi^2(11, n=1,852) = 40.36, p<0.001$ ).

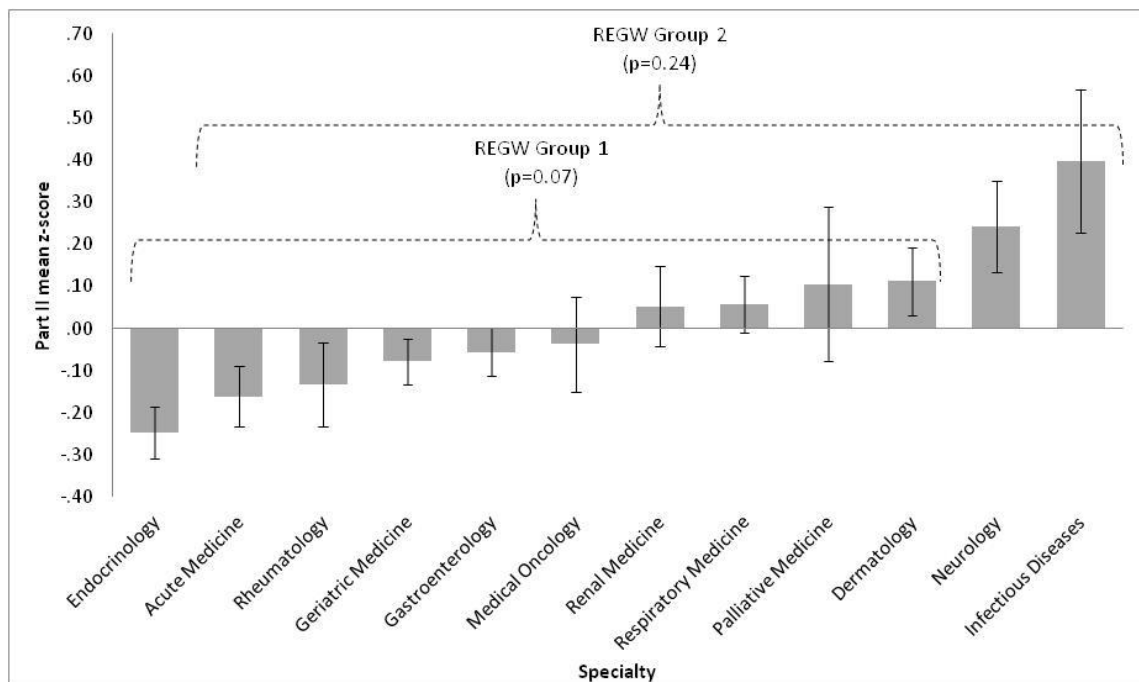


Figure 18. Comparison of mean Part II Z-scores (with SE) between specialties.

Inspection of Figure 18 indicates that candidates taking the SCE in Infectious Diseases and Neurology had significantly higher Part II scores than those taking other SCEs; however, the standard errors of those means were also quite high due to the small sample sizes. Candidates attempting Endocrinology had significantly lower Part II scores.

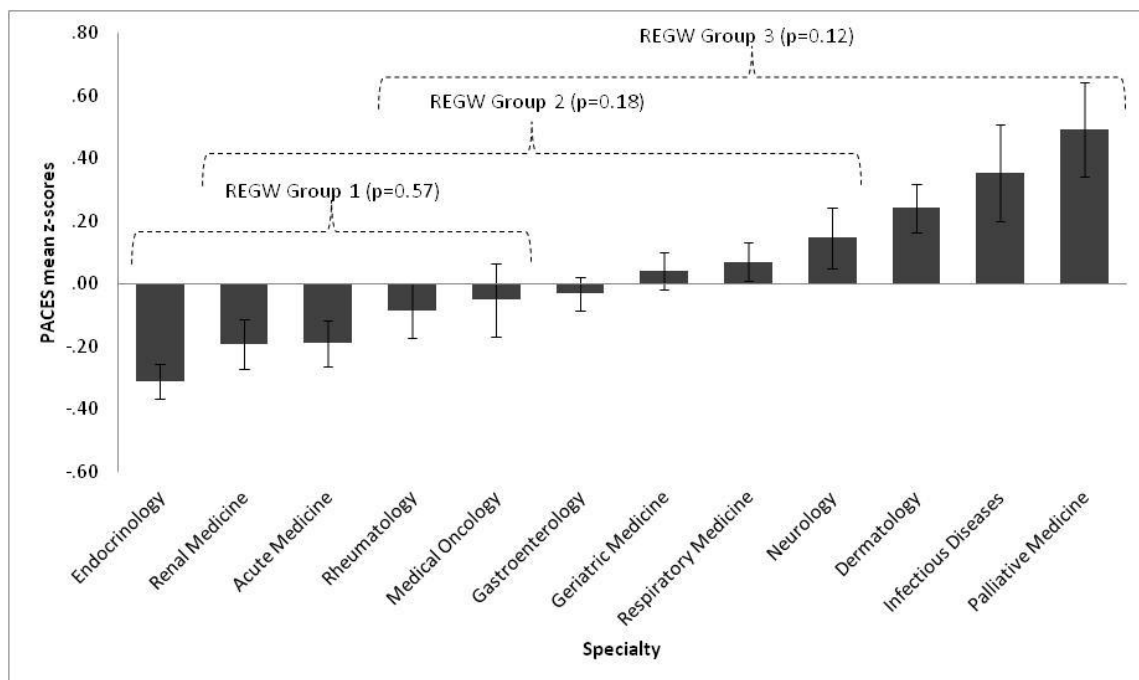


Figure 19. Comparison of mean PACES Z-scores (with SE) between specialties.

Similarly inspection of Figure 19 shows that candidates taking Palliative Medicine, Infectious Diseases and Dermatology exams had significantly higher PACES scores than the rest of the candidates, while candidates attempting Endocrinology had significantly lower PACES scores than others.

### **5.1.2 Inferential statistics**

#### **5.1.2.1 Correlations**

The values of the correlation coefficients between the first attempt scores in MRCP(UK) parts scores and the first attempt SCEs scores are presented in Table 29 (next page).

Inspection of Table 29 shows that Part II correlated most highly with the majority of the SCEs. The exceptions were the exams for Gastroenterology, Medical Oncology, Respiratory Medicine, Infectious Diseases, and Neurology exams, where the correlation coefficients with Part I were higher than or almost equal to those of Part II. The highest coefficient values were observed for Infectious Diseases with Part II ( $r=0.68$ ) and Part I ( $r=0.66$ ). The relationships between the SCEs and PACES were generally weaker in comparison to Part I and Part II, with the minimum value obtained for Endocrinology ( $r=0.16$ ). However, a relatively strong relationship with PACES was observed for Neurology, Palliative Medicine, Rheumatology, Infectious Diseases, and Respiratory Medicine.

The mean weighted correlation coefficients for Part II reached  $r=0.54$ , for Part I reached  $r=0.47$ , and in the case of PACES it was  $r=0.34$ .

**Table 29. Correlation coefficients (Pearson  $r$ ) between MRCP(UK) parts scores and SCE scores.**

<i>Specialty</i>	<i>Part I 1<sup>st</sup> Attempt</i>		<i>Part II 1<sup>st</sup> Attempt</i>		<i>PACES 1<sup>st</sup> Attempt</i>	
Acute Medicine	0.37**	$n=159$	0.48**	$n=191$	0.39**	$n=199$
Dermatology	0.48**	$n=125$	0.52**	$n=140$	0.36**	$n=150$
Endocrinology	0.29**	$n=147$	0.42**	$n=225$	0.16*	$n=270$
Gastroenterology	0.56**	$n=235$	0.53**	$n=298$	0.31**	$n=331$
Geriatric Medicine	0.49**	$n=233$	0.54**	$n=282$	0.35**	$n=296$
Infectious Diseases	0.66**	$n=34$	0.68**	$n=39$	0.43**	$n=43$
Neurology	0.62**	$n=63$	0.62**	$n=97$	0.50**	$n=126$
Medical Oncology	0.55**	$n=65$	0.48**	$n=72$	0.28*	$n=83$
Palliative Medicine	0.45*	$n=29$	0.66**	$n=31$	0.48**	$n=32$
Renal Medicine	0.45**	$n=103$	0.60**	$n=132$	0.28**	$n=153$
Respiratory Medicine	0.56**	$n=163$	0.49**	$n=220$	0.41**	$n=235$
Rheumatology	0.42**	$n=105$	0.63**	$n=125$	0.47**	$n=137$
<b>Mean (weighted) coefficient</b>	<b>0.47 (SD=0.10)</b>		<b>0.54 (SD=0.08)</b>		<b>0.34 (SD=0.09)</b>	

Significant \*\*  $p<0.01$ , \*  $p<0.05$

### ***Correction of correlation coefficients for range restriction and attenuation***

The coefficients presented in Table 29 were corrected for range restriction and disattenuated in accordance with the procedure described in Chapter 3, section 3.7.2. The correlation coefficients for each pair of an SCE and MRCP(UK) part were corrected separately. It was not taken into account that Part II and PACES were first restricted by previous MRCP(UK) parts. The standard deviations for unrestricted samples were the standard deviations of the first attempt scores in each of the MRCP(UK) parts for all candidates in the History File, while the restricted sample standard deviations were obtained from the records of those candidates who attempted an SCE only.

Table 30 (next pages) contains initial correlation coefficients, the reliability coefficients, standard deviations of the measures for the incumbent and unrestricted samples, as required by the range restriction procedure, and the corrected correlation coefficients.

The results presented in Table 30 show that the correction procedures resulted in an increase in the coefficients' magnitude by an overall 41% for Part I, 30% for Part II, and 27% for PACES, making the correlation coefficients very high. The corrected coefficients ranged from 0.46 to 0.86 for Part I ( $M=0.68$ ,  $SD=0.12$ ), 0.58 to 0.85 for Part II ( $M=0.72$ ,  $SD=0.08$ ), and 0.21 to 0.68 for PACES ( $M=0.46$ ,  $SD=0.12$ ); however, it needs to be noted that correction for range restriction and disattenuation is only an approximation of the true strength of the relationships.

**Table 30. Correlation coefficients between MRCP(UK) results and SCE results, by SCE, for each part of MRCP(UK).**

<i>Specialty</i>	<i>Correlation coefficient <math>r</math></i>	<i>Mean Reliability of Criterion</i>	<i><math>r</math> corrected for attenuation of Criterion</i>	<i>SD of a Part Results in Restricted Sample</i>	<i><math>r</math> corrected for Range Restriction</i>	<i>Final corrected <math>r</math></i>	<i>% change in <math>r</math></i>
<b>Part I (mean reliability = 0.91, SD* = 11.98)</b>							
Acute Medicine	0.37	0.83	0.41	8.27	0.54	0.57	53%
Dermatology	0.48	0.87	0.52	7.90	0.68	0.71	47%
Endocrinology	0.29	0.89	0.31	7.94	0.44	0.46	58%
Gastroenterology	0.56	0.82	0.62	9.24	0.72	0.75	34%
Geriatric Medicine	0.49	0.76	0.56	8.63	0.68	0.71	47%
Infectious Diseases	0.66	0.94	0.68	7.85	0.82	0.86	30%
Neurology	0.62	0.90	0.65	8.10	0.79	0.82	33%
Medical Oncology	0.55	0.83	0.61	9.53	0.69	0.72	32%
Palliative Medicine	0.45	0.82	0.50	7.57	0.67	0.70	56%
Renal Medicine	0.45	0.85	0.49	9.37	0.58	0.61	35%
Respiratory Medicine	0.56	0.84	0.62	9.02	0.72	0.75	34%
Rheumatology	0.42	0.90	0.44	9.72	0.52	0.54	30%
<b>MEAN: 0.49 (SD =0.10)</b>						<b>MEAN: 0.68 (SD=0.12)</b>	<b>MEAN: 41%</b>
<b>Part II (mean reliability =0.81, SD* = 7.62)</b>							
Acute Medicine	0.48	0.83	0.52	6.71	0.57	0.64	33%
Dermatology	0.52	0.87	0.56	6.41	0.63	0.70	33%
Endocrinology	0.42	0.89	0.45	6.20	0.53	0.58	38%
Gastroenterology	0.53	0.82	0.58	6.68	0.63	0.70	34%
Geriatric Medicine	0.54	0.76	0.62	6.10	0.70	0.78	44%
Infectious Diseases	0.68	0.94	0.70	7.12	0.73	0.81	19%

<i>Specialty</i>	<i>Correlation coefficient r</i>	<i>Mean Reliability of Criterion</i>	<i>r corrected for attenuation of Criterion</i>	<i>SD of a Part Results in Restricted Sample</i>	<i>r corrected for Range Restriction</i>	<i>Final corrected r</i>	<i>% change in r</i>
Neurology	0.62	0.90	0.65	7.23	0.67	0.75	21%
Medical Oncology	0.48	0.83	0.53	6.42	0.60	0.67	37%
Palliative Medicine	0.66	0.82	0.73	6.86	0.76	0.85	29%
Renal Medicine	0.60	0.85	0.65	7.36	0.66	0.73	23%
Respiratory Medicine	0.49	0.84	0.54	6.78	0.58	0.65	31%
Rheumatology	0.63	0.90	0.66	7.40	0.67	0.75	19%
MEAN:	0.55 (SD=0.08)					MEAN: 0.72 (SD=0.08)	MEAN: 30%
			<u>PACES (mean reliability =0.82, SD* = 6.93)</u>				
Acute Medicine	0.39	0.83	0.42	6.78	0.43	0.48	24%
Dermatology	0.36	0.87	0.38	6.25	0.42	0.47	30%
Endocrinology	0.16	0.89	0.17	6.02	0.19	0.21	35%
Gastroenterology	0.31	0.82	0.35	6.52	0.37	0.41	30%
Geriatric Medicine	0.35	0.76	0.40	6.66	0.42	0.46	32%
Infectious Diseases	0.43	0.94	0.45	6.65	0.46	0.51	19%
Neurology	0.50	0.90	0.53	7.12	0.52	0.57	15%
Medical Oncology	0.28	0.83	0.31	6.99	0.30	0.34	21%
Palliative Medicine	0.48	0.82	0.53	5.61	0.61	0.68	41%
Renal Medicine	0.28	0.85	0.30	6.49	0.32	0.35	28%
Respiratory Medicine	0.41	0.84	0.45	6.33	0.48	0.53	30%
Rheumatology	0.47	0.90	0.49	6.82	0.50	0.56	19%
MEAN:	0.37 (SD=0.10)					MEAN: 0.46 (SD=0.12)	MEAN: 27%

\* Standard deviation of scores for the unrestricted sample

### 5.1.2.2 Contrasting groups

Among the SCE candidates there were 514 MRCP(UK) Highfliers whose SCE scores could be compared with those of the other candidates. The comparison was made on Z-transformed SCE scores. Table 31 (below) presents a detailed breakdown of MRCP(UK) Highfliers counts by specialty, together with the results of the independent samples t-test for the differences in the standardised SCE scores between the Highfliers and Typical candidates (all others).

**Table 31. Comparison of mean scores between MRCP(UK) Highfliers and Typical Candidates for twelve specialties (independent samples t-test results) with effect sizes (*r*).**

<i>Specialty</i>	<i>N Typical</i>	<i>N Highfliers</i>	<i>Typical Candidates</i>	<i>MRCP(UK) Highfliers</i>	<i>Independent samples t- test result</i>	<i>Effect sizes (r)</i>
Acute Medicine	177	38	-0.07	0.79	$t(213) = -4.89, p < 0.001$	0.32
Dermatology	103	60	-0.18	0.51	$t(161) = -4.82, p < 0.001$	0.36
Endocrinology	257	36	0.02	0.48	$t(291) = -2.54, p < 0.001$	0.15
Gastroenterology	270	74	-0.17	0.73	$t(342) = -7.17, p < 0.001$	0.36
Geriatric Medicine	225	79	-0.16	0.75	$t(302) = -7.77, p < 0.001$	0.41
Infectious Diseases	27	20	-0.28	0.66	$t(45) = -3.96, p < 0.001$	0.22
Neurology	118	31	-0.02	0.58	$t(147) = -3.15, p < 0.001$	0.25
Medical Oncology	83	21	-0.26	1.00	$t(102) = -5.60, p < 0.001$	0.48
Palliative Medicine	26	12	-0.27	0.59	$t(36) = -2.67, p = 0.016$	0.41
Renal Medicine	131	37	-0.06	0.68	$t(166) = -4.18, p < 0.001$	0.31
Respiratory Medicine	184	62	-0.22	0.77	$t(244) = -7.17, p < 0.001$	0.42
Rheumatology	101	44	-0.22	0.76	$t(143) = -6.35, p < 0.001$	0.47
<b>Total:</b>	<b>1,702</b>	<b>514</b>	<b><math>M_W = -0.12^*</math> (SD=0.03)</b>	<b><math>M_W = 0.69^*</math> (SD=0.17)</b>	$t(2,423) = -17.90, p < 0.001$	0.34

\*  $M_W$ : weighted means

The contents of Table 31 shows that MRCP(UK) Highfliers scored significantly higher in all SCEs in comparison to the rest of the candidates; the mean effect size estimated with Cohen's *d* was 0.73 ( $r=0.34$ ), which is considered moderate.



### 5.1.2.3 Regression models

Twelve linear regression models were fitted for each specialty separately using the entry method to assess the joint effect of the MRCP(UK) parts scores on the standardised SCE scores. Table 32 summarizes the models.

**Table 32. Summary of the regression models fitted for the SCE scores with MRCP(UK) parts as predictors.**

<i>Specialty</i>	<i>Standardised coefficients for</i>			<i>R<sup>2</sup></i>	<i>Average VIF</i>
	<i>Part I</i>	<i>Part II</i>	<i>PACES</i>		
Acute Medicine model ( <i>n</i> =154)	0.12	0.36**	0.28**	0.339	1.40
Dermatology model ( <i>n</i> =125)	0.29*	0.31*	0.04	0.301	1.45
Endocrinology model ( <i>n</i> =146)	0.14	0.33**	0.15*	0.214	1.17
Gastroenterology model ( <i>n</i> =233)	0.38**	0.25**	0.12*	0.384	1.48
Geriatric Medicine model ( <i>n</i> =233)	0.26**	0.35**	0.16*	0.373	1.33
Infectious Diseases model ( <i>n</i> =34)	0.25	0.54*	0.06	0.606	1.94
Neurology model ( <i>n</i> =63)	0.31*	0.38*	0.16	0.492	1.52
Medical Oncology model ( <i>n</i> =62)	0.40*	0.13	0.09	0.291	1.71
Palliative Medicine model ( <i>n</i> =29)	0.06	0.51 <sup>(a)</sup>	0.21	0.482	2.32
Renal Medicine model ( <i>n</i> =102)	0.11	0.49**	-0.04	0.314	1.69
Respiratory Medicine model ( <i>n</i> =164)	0.33**	0.36**	0.06	0.394	1.33
Rheumatology model ( <i>n</i> =105)	0.12	0.42**	0.22*	0.394	1.44

\*\* significant with  $p < 0.001$ , \* significant with  $p < 0.05$ , (a) –almost significant at  $p = 0.054$

All SCE models were tested for assumptions; hence, Table 32 also contains the values of the VIF statistic (see Chapter 3, section 3.7.4). Multicollinearity was not observed in any of the models as indicated by the values of the average VIF.

Results gathered in Table 32 show that Part II scores were the best predictor for the majority of specialties, as was previously observed in the size of the correlation coefficients. Also in the case of the regression models, several exceptions were encountered: in the cases of Gastroenterology and Medical Oncology, Part I scores were the best predictor, while for Neurology, Dermatology, and Respiratory Medicine, both written parts had a similar impact. PACES did not explain any variance in the SCE performance for the majority of specialties after taking into account Part I and Part II scores. However, the exceptions

were Acute Medicine, Endocrinology, Gastroenterology, Geriatric Medicine, and Rheumatology.

Interestingly, in the case of Acute Medicine, Endocrinology, Infectious Diseases, Renal Medicine, Rheumatology, and Palliative Medicine, where the SCEs scores were quite strongly correlated with Part I scores, the regression models yielded a non-significant Part I beta coefficient. In the case of Medical Oncology, Part II scores became a non-significant predictor despite previously obtained relatively high correlation coefficient. These observations could be explained by quite strong correlation between Part I and Part II scores, and the regression models assigning variability of the SCEs scores to just one of those measures; the one where the association was stronger.

### ***Correction of regression models coefficients for range restriction***

To correct the linear models for range restriction, the EM algorithm was employed on Z-transformed MRCP(UK) scores so as to find the missing values, which constituted approximately from 17% to 36% of the sample (for justification see Chapter 3, section 3.7.2). The EM algorithm implementation resulted in an almost negligible shift in distributions of the MRCP(UK) means. The provided estimates were used to fit the new regression models for each SCE separately using the entry method. The summary for those models is presented in Table 33 (next page).

The pattern of significance for the predictors did not change in comparison to previous models that used data with missing values (Table 32). The largest observed difference was that PACES became a significant predictor for Respiratory Medicine scores. At the same time it became significant in predicting Neurology scores in place of Part I scores and stopped being predictive of Endocrinology scores. Further, an observed change was that predictors generally became smaller in magnitude. The new models also explained less variance in comparison to the original models based on *R*-squared values, and the multicollinearity increased based on the average VIF. These results were not surprising, given that the EM algorithm uses correlation to estimate the missing values. With a large amount of missing data, the algorithm tends to replace the missing data with values varying around means.

**Table 33. Summary of the regression models fitted for the SCE scores with missing MRCP(UK) parts scores estimated with the EM algorithm.**

<i>Specialty</i>	<i>Standardised coefficients for</i>			$R^2$	<i>Average VIF</i>
	<i>Part I</i>	<i>Part II</i>	<i>PACES</i>		
Acute Medicine model ( $n=225$ )	0.06	0.28*	0.20*	0.180	1.49
Dermatology model ( $n=165$ )	0.26*	0.27*	0.12	0.282	1.55
Endocrinology model ( $n=295$ )	0.11	0.27**	0.03	0.128	1.36
Gastroenterology model ( $n=348$ )	0.32**	0.25**	0.11*	0.331	1.56
Geriatric Medicine model ( $n=305$ )	0.22**	0.32**	0.15*	0.317	1.46
Infectious Diseases model ( $n=49$ )	0.26	0.46*	~0.00	0.451	2.09
Neurology model ( $n=150$ )	0.19	0.25*	0.25*	0.341	2.05
Medical Oncology model ( $n=105$ )	0.34*	0.19	0.04	0.258	1.76
Renal Medicine model ( $n=173$ )	0.07	0.46**	0.04	0.285	1.95
Respiratory Medicine model ( $n=246$ )	0.25**	0.25**	0.24**	0.335	1.49
Rheumatology model ( $n=145$ )	0.11	0.41**	0.19*	0.365	1.56
Palliative Medicine model ( $n=38$ )	0.17	0.26	0.32 <sup>(a)</sup>	0.378	2.15

\*\* significant with  $p<0.001$ , \* significant with  $p<0.05$ , (a) almost significant at  $p=0.061$

#### **5.1.2.4 Similarity of the models**

The fitted regression models have shown that MRCP(UK) scores predicted the results of the SCEs and their influence was cumulative, which is visualised in Figure 20. The regression lines represent the twelve specialties under investigation. The zeros on the axes denote the mean or the pass-mark, while units represent standard deviations. The dotted lines are added for ease of reference. The lines on Figure 20 are almost parallel and therefore a question arose as to whether there were any significant differences between the twelve models.

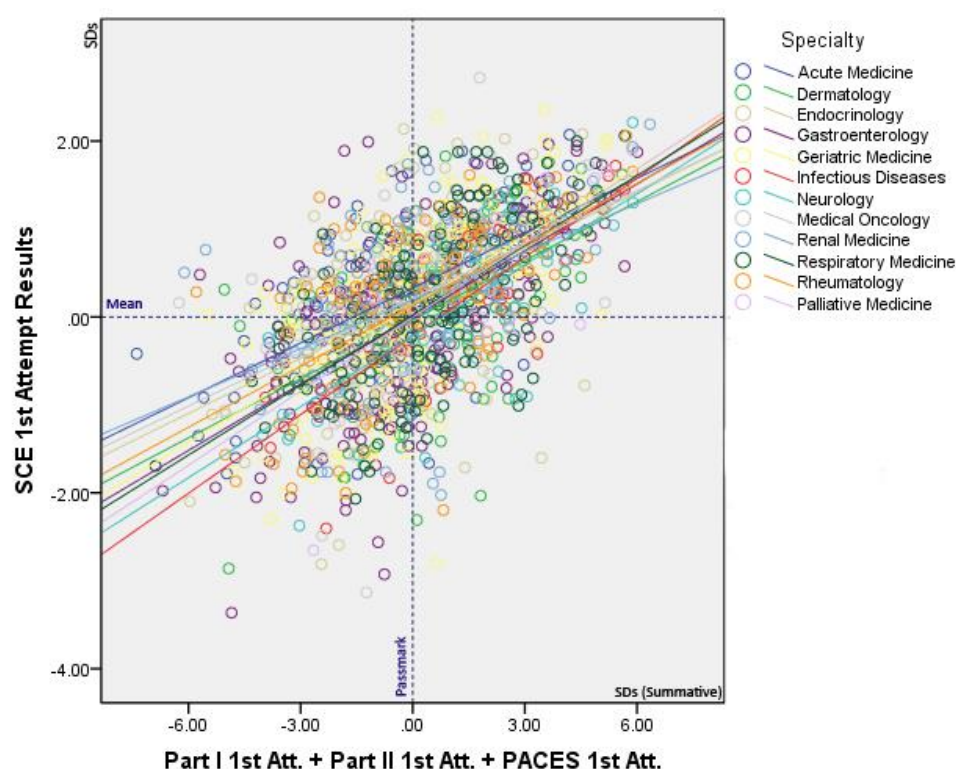


Figure 20. Fitted separate regression lines for SCEs with aggregated MRCP(UK) Z-scores.

In order to investigate this issue, the differences between the slopes of the regression lines were tested for significance using multilevel modelling. Use of the multilevel model was appropriate as the SCEs data could be treated as nested; an alternative approach would require multiple pair-wise comparisons using Chow tests, which was not feasible.

Figure 21 (next page) presents the output obtained from the MLwiN software (see Section 3.8, Chapter 3) with the general SCE multilevel model. It is represented by the equation, in which the estimates of  $\beta_{0ij}$  to  $\beta_{3j}$  are associated with the standardised first attempt Part I scores, Part II scores, and PACES scores. The covariance matrix (indicated as  $u$  terms) is provided underneath. Values on diagonal of that matrix indicate the variability within each  $\beta$  term (for example, for Part I score  $\beta$  this is marked with a thick-lined blue oval). Should these terms be significant, it would indicate a high variability of the slopes. The covariance of the first two  $u$  terms marked with smaller fine-lined oval represents the relationship between the intercept and the slope of the regression lines as if it the model comprised only beta coefficient for Part I. Analogously, other covariance terms represent the dependency of beta coefficients from other beta coefficients (the larger black-lined oval labelled; marked as “covariance terms”).

Equations

$$Zpercent_{ij} \sim N(XB, \Omega)$$

$$Zpercent_{ij} = \beta_{0ij}constant + \beta_{1j}ZPart1MarkAttempt1_{ij} + \beta_{2j}ZPart2MarkAttempt1_{ij} + \beta_{3j}ZPACESMarkAttempt1_{ij}$$

$$\beta_{0ij} = 0.091(0.039) + u_{0j} + e_{0ij}$$

$$\beta_{1j} = 0.244(0.032) + u_{1j}$$

$$\beta_{2j} = 0.325(0.025) + u_{2j}$$

$$\beta_{3j} = 0.129(0.022) + u_{3j}$$

← Estimates for the general model

$$\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.012(0.007) & -0.009(0.005) & 0.004(0.004) & 0.000(0.000) \\ 0.000(0.000) & 0.000(0.000) & 0.000(0.000) & 0.000(0.000) \\ 0.000(0.000) & 0.000(0.000) & 0.000(0.000) & 0.000(0.000) \\ 0.000(0.000) & 0.000(0.000) & 0.000(0.000) & 0.000(0.000) \end{bmatrix}$$

Figure 21. Summary of the multi-level model for the SCE scores with MRCP(UK) parts as predictors (MLwiN output).

The numbers in the output represent the estimates of beta coefficients for the model (rounded rectangle). For example, the number 0.244 in the upper left corner ( $\beta_{1j}$ ) shows that an increase by one standard deviation in Part I scores results in a 0.244 increase in an SCE score. Analogously, the values of 0.325 and 0.129 should be interpreted as representing an increase in SCE scores resulting from one standard deviation increase in Part II and PACES scores. These beta coefficients were statistically significant, as indicated by the values of the associated error terms (provided in brackets), which are much smaller than the values of the coefficients. For example, the confidence intervals for  $\beta_{1j}$  were estimated as  $\beta_{1j} \pm 1.96 SE$ , which provided 95%CI [0.181, 0.307], which did not encompass zero (McHugh, 2008).

The  $u$ -terms matrix lacked statistically significant variance and covariance terms, as also indicated by the values of the  $u$ -terms and their associated errors. This suggests that there was hardly any variability within the slopes of the regression lines, and that there was no relationship between the intercept and the slope in that model; both of which would signify the regression lines crossing one another. Therefore, based on the above results, it was assumed that the regression lines of different specialties were almost parallel, as seen in Figure 20.

Following this finding, the SCE general linear regression model was fitted using all available Z-transformed first attempt results across specialties ( $n= 1,449$ ). As presented in Table 34 the new model explained 32.6% of variance and was considered moderately well-fitted. For

the purposes of correction for range restriction the model was also re-fitted using EM estimated data, which increased the number of valid cases by 767 (model marked with an asterisk: \*). Both models are presented in Table 34 (below).

**Table 34. Summary of the general SCEs linear model with and without EM algorithm substituted missing values.**

<i>Dependent variable / Model</i>	<i>Standardised coefficients for</i>			<i>R<sup>2</sup></i>	<i>Average VIF</i>
	<i>Part I</i>	<i>Part II</i>	<i>PACES</i>		
General SCEs model	0.24**	0.34**	0.12**	0.326	1.41
General SCEs model*	0.20**	0.29**	0.13**	0.262	1.56

\*\* significant at  $p < 0.001$ ,

The model with EM algorithm imputed missing values explained less variance and the predictors were weaker in comparison to the model based on the raw data, as was previously observed for individual SCEs. However, both models showed that all three parts of MRCP(UK) had a significant effect on predicting specialty exam scores.

## 5.2 CARDIOLOGY CKBA

Cardiology Knowledge Based Assessment is the final examination in the Cardiology specialty. CKBA is administered by the British Cardiovascular Society and consists of 120 questions. It tests for the use of knowledge in a clinical setting, and therefore, it is similar to the Part II examination. More details on the exam can be found in section 2.4.5.

### 5.2.1 Descriptive statistics

The sample consisted of 209 doctors of whom 31 (14.8%) were female and 178 were male (85.2%). Most of them were non-white doctors (61.2%) who qualified (71.8%) and trained (99%) in the United Kingdom. The distribution of the CKBA scores in this sample was close to normal, as indicated by the value and significance of the K-S test ( $Z_{KS} = 1.08$ ,  $p = 0.33$ ). The results ranged from 45.83 to 84.75 percentage points with mean equal to 69.79 (SD=7.03).

The distributions of the first attempt scores in Part I, Part II, and PACES in this sample were also tested for normality with the K-S test, which indicated that they were close to normal.

## 5.2.2 Inferential Statistics

### 5.2.2.1 Correlations

Scores in the MRCP(UK) parts and the Cardiology CKBA showed reasonably high correlations (Part I  $r=0.53$ ,  $p<0.001$ ; Part II  $r=0.51$ ,  $p<0.001$ ; PACES  $r=0.34$ ,  $p<0.001$ ). Table 35 (below) provides the values of the standard deviations for restricted and unrestricted samples and the reliability coefficients required to apply the correction for range restriction and disattenuation. As explained previously in section 5.1, the unrestricted sample SDs were the standard deviation of first attempt scores in MRCP(UK) parts for all candidates present in the History File, while restricted sample SDs were obtained for those candidates who attempted CKBA. The correction did not take into account that Part II and PACES were also restricted by Part I, or Part I and Part II, respectively.

**Table 35. Parameters for derestriction of range and disattenuation of coefficients between CKBA and MRCP(UK).**

<i>Measure</i>	<i>SD Restricted Sample</i>	<i>SD Unrestricted Sample</i>	<i>Reliability</i>
Part I	8.32	11.98	0.91
Part II	6.16	7.62	0.81
PACES	5.92	6.93	0.82
Cardiology	n/a	n/a	0.75

Application of the range restriction correction and disattenuation resulted in an increase in those coefficients to  $r=0.77$  (increase by 48%),  $r=0.75$  (increase by 46%), and  $r=0.49$  (increase by 46%).

### 5.2.2.2 Contrasting groups

In the sample of 209 Cardiology candidates there were 68 MRCP(UK) Highfliers. A significant difference in the CKBA results was observed between the Highfliers ( $M=0.46$ ;  $z$  transformed scores) and the rest of the candidates ( $M=-0.22$ ) based on the value of the independent samples  $t$ -test ( $t(207)=-4.91$ ,  $p<0.001$ ,  $r=0.32$ ). As in the case of previous specialty exams, the MRCP(UK) Highfliers scored higher.

### 5.2.2.3 Regression models

The MRCP(UK) first attempt  $Z$ -transformed scores were regressed onto CKBA first attempt  $Z$ -transformed scores using the entry method. The model was fitted based on 123 cases and explained 36.8% of variance. The standardised beta coefficients equalled 0.30 ( $p=0.001$ ) for

Part I, 0.26 ( $p=0.005$ ) for Part II, and 0.20 ( $p=0.010$ ) for PACES, suggesting that Part I was the best predictor. However, the coefficients for Part II and PACES were only slightly smaller. Analysis of the part and partial correlations for this model showed that Part I, Part II, and PACES had very similar predictive values.

#### ***Correction of regression coefficients for range restriction***

Similar to the other SCE models, an EM algorithm was applied to maximise the sample size ( $n=209$ ). The new model was fitted using the entry method and it explained 29% of variance. The beta coefficients reached 0.25 ( $p=0.002$ ) for Part I, 0.24 ( $p=0.004$ ) for Part II, and 0.16 ( $p=0.015$ ) for PACES. As was the case for the other SCE models, the model based on EM algorithm imputed data explained less variance and the beta coefficients were smaller.

#### ***5.2.2.4 Similarity of the Cardiology model to other SCEs***

The CKBA model was compared with other SCE models using multi-level modelling. CKBA results were merged with the SCE data and a new joint model was calculated with MLwiN for all thirteen specialties. The output analogous to the previous one is presented in Figure 22.

The important parameters of this model, which were the variances and covariances ( $u$ -terms), did not differ from the general SCE model (Figure 21) and the interpretation remained unchanged. The effect of the intercept in this model, despite being seemingly statistically significant, was not. Expanding the values from the output to four decimal places, it was found that the standard error was 0.0095 while the effect was 0.0182;  $1.96 \times 0.0095 = 0.0186$  was more than 0.0182, and therefore, it was likely that effect could have assumed the value of zero (McHugh, 2008). A similar observation was made with the covariance between the intercept and  $\beta_{1j}$ , where the calculated value for  $SE$  and the effect were approximately equal. Based on the values in the covariance matrix, there was no interaction between the intercept and the slopes of the lines. Therefore, the CKBA model was assumed to be similar to the other SCE models.



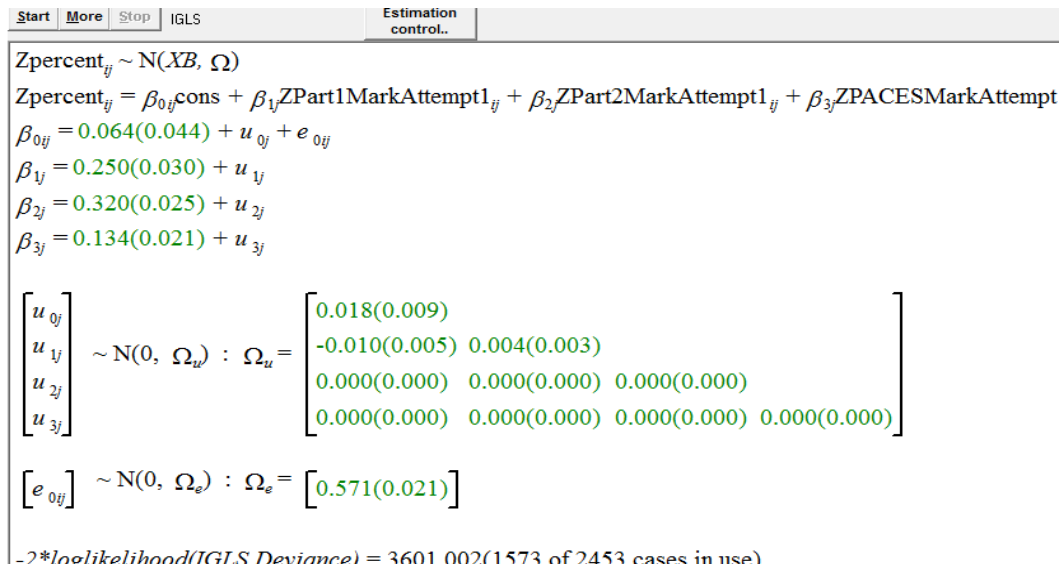


Figure 22. Summary of the multi-level model for the SCEs with Cardiology with MRCP(UK) parts as predictors (MLwiN output).

Based on this finding, a general model for thirteen specialties was fitted ( $n=1,573$ ) and corrected for range restriction ( $n=2,453$ ). The entry method was employed. Both models are summarised in Table 36 (below).

**Table 36. Summary of the general SCEs & Cardiology linear model with missing values estimated using EM algorithm.**

Model	Standardised coefficients for			$R^2$	Average VIF
	Part I	Part II	PACES		
SCEs model (with Cardio)	0.24**	0.32**	0.12**	0.301	1.43
SCEs model (with Cardio)*	0.19**	0.28**	0.12**	0.234	1.58

\*\* significant at  $p < 0.001$

Data in which missing values were replaced with the EM algorithm estimates yielded a less well-fitted and generally weaker model. The SCE model including CKBA was very similar to the model obtained for the SCEs only. This general model was further used for comparison purposes with regression models fitted for other exams.

### 5.3 FRCR – CLINICAL ONCOLOGY

The examination for the Fellowship of the Royal College of Radiologists ('FRCR') is a complex exam consisting of two parts: the First examination ('FRCR1') and Final Examination ('FRCR2'). The FRCR1 exam divides into four modules that can be taken over a period of two years. It tests for theoretical and practical knowledge in Statistics, Physics,

Cancer Biology and Radiobiology, and Clinical Pharmacology. As such FRCR1 was assumed to be representative to the construct of medical knowledge. The FRCR2 examination comprises written, oral, and clinical components. Whereas written and oral components were assumed to be measures of knowledge with an aspect of communication skills in the case of the oral exam, the clinical exam was assumed to be a measure of purely clinical skills, and as such it is analysed in Chapter 6. More details on the FRCR exam can be found in section 2.4.3.

### **5.3.1 Descriptive statistics**

#### **5.3.1.1 FRCR1**

The analysed file contained records of 756 individuals, but only 746 records were complete. Additionally, two records were coded with an error and were removed from the dataset. Among 744 doctors on the dataset 350 (47.0%) were male and 392 (52.7%) were female (2 cases missing). Of 474 candidates with a full demographic record 311 (65.6%) declared other than white ethnicity, 352 (74.3%) were UK graduates, and 448 (94.5%) were UK trainees. Based on the FRCR1 Rank in the provided dataset, there were 202 Highfliers, 450 Typical candidates, 20 Dropouts, and 26 Failed candidates. Additionally, there were 46 individuals with censored data.

As previously described in Chapter 2, section 2.4.3, the FRCR1 comprises four modules. Table 37 presents a breakdown of the numbers of candidates who did not attempt a particular module and who passed and failed a particular module. It also presents a breakdown of the number of attempts at which a module was passed and the overall passing rate for each module. Candidates considered ‘censored’ – meaning they did not have a chance to attempt all modules at the time when the datasets were provided for this research, and therefore their ultimate score was not known – were excluded.

**Table 37. Number of candidates who passed or failed FRCR1 modules with the number of attempts and pass-rates.**

<i>Candidates who:</i>	<i>Cancer Biology (n=698)</i>	<i>Clinical Pharmacology (n=694)</i>	<i>Medical Statistics (n=692)</i>	<i>Physics (n=693)</i>
<b>Not tried so far</b>	<b>0</b>	<b>4</b>	<b>6</b>	<b>5</b>
<b>Failed (total)</b>	<b>29</b>	<b>17</b>	<b>24</b>	<b>29</b>
<b>Passed</b>	<b>669</b>	<b>677</b>	<b>668</b>	<b>664</b>
on 1 <sup>st</sup> attempt	487	550	498	491
on 2 <sup>nd</sup> attempt	143	92	120	134
on 3 <sup>rd</sup> attempt	33	25	41	29
on 4 <sup>th</sup> attempt	6	10	8	9
on 6 <sup>th</sup> attempt	0	0	1	1
<b>Mean Number of Attempts</b>	<b>1.37</b>	<b>1.27</b>	<b>1.39</b>	<b>1.38</b>
<b>Pass- rate (Passed versus Total)</b>	<b>95.8%</b>	<b>97.6%</b>	<b>96.5%</b>	<b>95.8%</b>

Note: 6 attempts was an exception (one person); this person's data were included in the analyses as their Total Number of Attempts in FRCR1 was still within the assumed limits of attempts.

The small differences between the difficulty levels were not statistically significant.

The distributions of the Total Number of Attempts for all candidates in total and by FRCR1 Rank are presented in Table 38. The distributions were significantly different from normal ( $p < 0.001$ ).

The majority (652 of 744, 87.6%) of the candidates passed FRCR1. Of those, more than half passed FRCR1 (358 candidates) with four attempts, which means they had only one attempt in each module. Nearly a third of those who passed (202 of 652) passed all modules during one diet (the group of Highfliers), while a third of those who passed (131) did so in more than one diet. However, based on the information obtained from the RCR, the candidates are in fact encouraged to attempt all modules in one diet, which might partially explain the large number of Highfliers.

**Table 38. Counts of the Total Number of Attempts in FRCR1, overall and in division by the FRCR1 Rank groups.**

<i>Total No. of attempts</i>	<i>Censored N</i>	<i>Failed N</i>	<i>Dropouts N</i>	<i>Typical N</i>	<i>Highfliers N</i>	<i>Total N</i>
1	0	0	1	0	0	1
2	4	0	3	0	0	7
3	2	0	3	0	0	5
4	20	0	5	131	202	358
5	4	0	0	129	0	133
6	2	3	3	74	0	82
7	8	2	1	46	0	57
8	4	3	0	35	0	42
9	1	3	1	14	0	19
10	1	2	2	9	0	14
11	0	4	0	3	0	7
12	0	4	1	4	0	9
13	0	2	0	1	0	3
14	0	2	0	4	0	6
15	0	1	0	0	0	1
<b>Total N</b>	<b>46</b>	<b>26</b>	<b>20</b>	<b>450</b>	<b>202</b>	<b>744</b>

Further inspection of Table 38 shows that there were 46 censored candidates, and 46 Dropouts and Failed candidates. Interestingly, those who dropped out did so relatively early in the process after having 1 to 6 attempts, while those who failed attempted each module several times. This was reflected in the mean and the median for the Total Number of Attempts in FRCR1 when examining the FRCR1 Rank groups, as presented in Table 39.

**Table 39. Mean and median scores for Total Number of Attempts in FRCR1 by FRCR1 Rank.**

<i>Total Number of Attempts in FRCR1</i>	<i>Failed</i>	<i>Dropouts</i>	<i>Typical</i>	<i>Highfliers</i>
Mean	10.15	5.10	5.73	4.0
Median	10.50	4.0	5.0	4.0
SD	2.64	3.09	1.89	0.00
n	26	20	450	202

The differences between the FRCR1 Rank groups in the Total Number of Attempts were statistically significant as indicated by the one-way ANOVA results ( $F(3,273)=15.71$ ,

$p < 0.001$ ). Due to heavily skewed distributions and lack of homogeneity of variance ( $F(3,273)=9.86$ ,  $p < 0.001$ ), bootstrapping was applied. A REGW Q test revealed that a significant difference was present between the candidates who failed FRCR1 (Failed:  $M=10.15$  95%CI [9.08, 11.22], bias=0.021,  $SE=0.50$ ) and the rest of the groups, with Drop-out candidates ( $M=5.10$  95%CI [3.65, 5.91], bias=-0.023,  $SE=0.68$ ) and Typical candidates ( $M=5.73$  95%CI [5.56, 5.91], bias=0.002,  $SE=0.09$ ) constituting a homogenous group ( $p=0.187$ ), and with Highfliers ( $M=4.0$ , bias=0.00,  $SE=0.0$ ) forming the last group.

Further, it was found that the groups of Dropouts and Failed candidates differed in terms of the number of modules attempted and modules passed, as presented in Table 40 (below).

**Table 40. Comparison of the number of modules attempted and passed for groups of Dropouts and Failed.**

	Group	None	1 module	2 modules	3 modules	4 modules	Total Count
Modules attempted - Candidate count	<b>Dropouts</b>	n/a	1	4	4	11	<b>20</b>
	<b>Failed</b>	n/a	0	0	0	26	<b>26</b>
Modules passed - Candidate count	<b>Dropouts</b>	13	4	3	0	0	<b>20</b>
	<b>Failed</b>	1	3	9	13	0	<b>26</b>

As presented in the table, in the group of those who failed FRCR1 all candidates attempted all four modules and half of them managed to pass three out of four (13 candidates). Only one person failed all modules after attempting all of them. On the other hand, in the group of Dropouts, only one person decided to drop out after attempting just one module, while most of them attempted all (11 candidates). Thirteen Dropouts failed all modules, which explains why they decided to cease their efforts.

Groups differentiated according to the FRCR1 Rank differed not only in the Total Number of Attempts, but also in the Mean Module Mark on 1<sup>st</sup> Attempt. The scores were normally distributed in the analysed sample of 277 candidates (K-S test,  $Z_{KS} = 0.64$ ,  $p=0.812$ ). Figure 23 presents the mean scores for the Mean Module Mark on 1<sup>st</sup> Attempt variable in four groups based on the FRCR1 Rank.

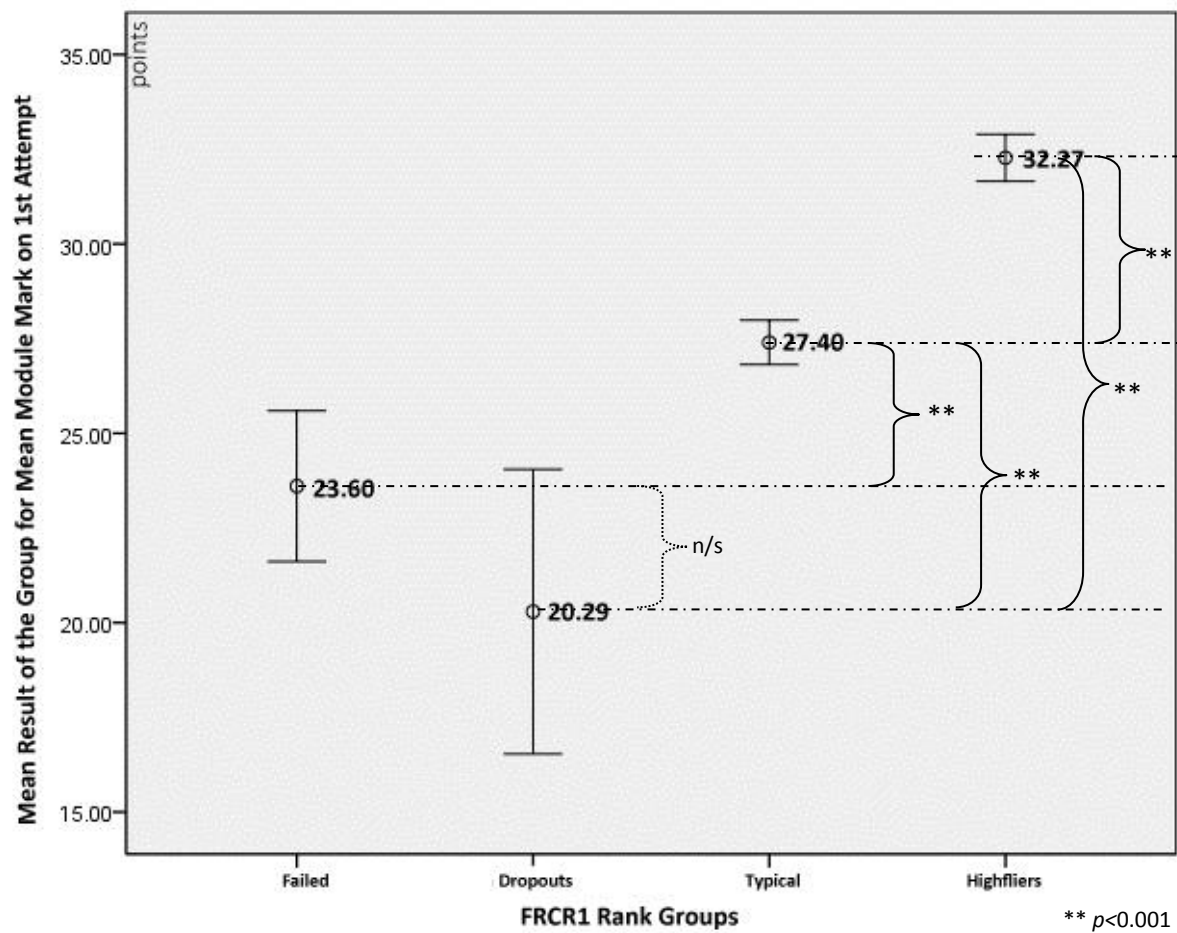


Figure 23. Comparison of first attempt Mean Module Marks (with 95% CI) between FRCR1 Rank groups.

Figure 23 shows that the Dropouts had the lowest scores, while Highfliers had the highest scores. A one-way ANOVA test showed the mean scores between groups to be highly significant ( $F(3,276)= 35.75$ ,  $p<0.001$ ); however, the variances in the groups were not homogenous, as indicated by the Levene's test ( $F(2,273)= 6.70$ ,  $p<0.001$ ), and therefore the bootstrapping method was employed. Post-hoc REGW Q tests showed that the groups of Failed ( $M=23.60$  95%CI [21.61, 25.59], bias=0.020,  $SE=0.90$ ) and Dropout ( $M=20.29$  95%CI [16.53, 24.05], bias=0.077,  $SE=0.48$ ) candidates constituted one homogenous group ( $p=0.144$ ), and the Typical candidates ( $M=27.40$  95%CI [27.66, 28.94], bias=0.023,  $SE=0.30$ ) and the FRCR1 Highfliers ( $M=32.27$  95%CI [31.65, 32.89], bias=0.004,  $SE= 0.30$ ) groups were separate homogenous groups. Based on these results, the Failed and Dropout groups were combined together in further analyses.

### 5.3.1.2 FRCR2

The dataset with FRCR2 results contained 337 records of candidates of which 159 (47.2%) were male and 178 (52.8%) were female, and of the 251 individuals with full records, 70.2% (177) declared non-white ethnicity, 82.9% (208) were UK graduates, and 98.4% (248) were UK Trainees.

Based on the provided data, the distribution of Total Number of Attempts in FRCR2 was not normal ( $Z_{KS} = 5.55$ ,  $p < 0.001$ ), which is also apparent from the data in Figure 24 (below). Inspection of Figure 24 shows that the majority of the candidates (180; 53% overall) passed FRCR2 on their first attempt and a further 22.6% (76 candidates) passed on their second attempt. The distribution was therefore highly skewed.

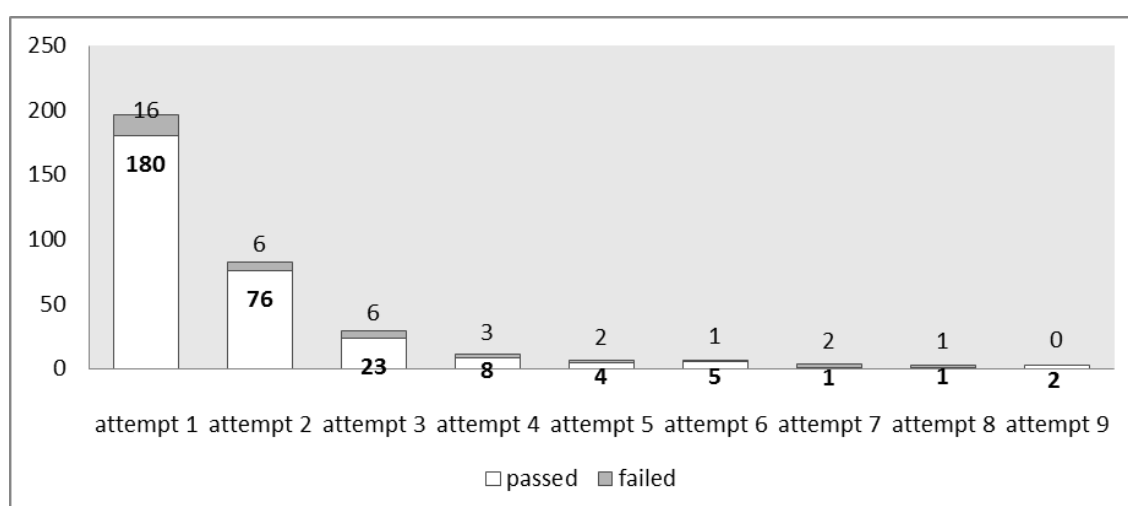


Figure 24. Distribution of the Total Number of Attempts in FRCR2 by candidates who passed and failed FRCR2.

As shown in Figure 24, the majority of the candidates who failed FRCR2 (22 of 37) had a record of only one or two attempts. Therefore, it was very likely that they would have passed the exam with further attempts. This means that at the time of performing the analyses, they may have been considered right-censored.

FRCR2 Rank, analogous to the FRCR1 Rank, was not proposed as it did not seem feasible. In contrast to FRCR1, which is a multi-module exam where each module is scored separately, FRCR2 consists of Part A and Part B, which are considered partial to an overall mark. The FRCR2 results were, however, analysed based on the FRCR1 Rank, but only for those candidates who passed FRCR1. The comparison was therefore effectively performed

between Typical candidates and Highfliers. The results of such a comparison are presented in Table 41.

**Table 41. Performance in FRCR2: mean scores (with SD) and mean number of attempts in division by FRCR1 Rank groups.**

<i>Variable</i>	<i>Typical candidates</i>			<i>Highfliers</i>		
	<i>Mean</i>	<i>SD</i>	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>N</i>
Number of attempts in FRCR2	2.02	1.50	<i>n</i> =232	1.38	1.00	<i>n</i> =105
FRCR2 Pass/Fail (1-pass)	0.86	0.34	<i>n</i> =232	0.94	0.23	<i>n</i> =105
1 <sup>st</sup> attempt Written (Z-score)	0.01	0.98	<i>n</i> =219	0.30	1.11	<i>n</i> =102
1 <sup>st</sup> attempt Oral (Z-score)	-0.01	0.95	<i>n</i> =218	0.41	0.92	<i>n</i> =102

Inspection of Table 41 shows that Highfliers needed fewer attempts to pass FRCR2 and on average they passed better (had fewer fails) than the group of Typical candidates. The difference between Typical candidates and Highfliers was statistically significant for the Total Number of Attempts in FRCR2 ( $t(335)=3.94$ ,  $p<0.001$ ,  $r=0.21$ ) and for the Pass/Fail Record ( $t(335)=-2.09$ ,  $p=0.038$ ,  $r=0.11$ ). Those who belonged to the Highfliers group were 2.54 times more likely (odds ratio) to pass the exam.

Table 41 also presents the mean standardised partial results in FRCR2, namely the written and oral components of FRCR2. The means for the Typical candidates' scores varied around 0.00 while the Highfliers scored approximately +0.3 to +0.4 SD. The differences were tested for significance using the independent samples t-test, as it was previously indicated that both parametric and non-parametric tests yield analogous results (see Chapter 4 for justification). The differences were statistically significant for the written component ( $t(318)=-2.35$ ,  $p=0.020$ ,  $r=0.13$ ) and for the oral component ( $t(318)=-3.66$ ,  $p<0.001$ ,  $r=0.20$ ).

The above-presented mean scores for both groups indicated that the mean of the overall FRCR2 sample was above zero, which may be surprising. However, there are two potential explanations for that phenomenon. Firstly, due to the fact that the raw results provided by the RCR were in different formats depending on the diet of the exam (see Chapter 3, section 3.2.3), the transformation to Z-scores was performed within each diet separately. The analysed data, on the other hand, were aggregate of those standardised results. Figure 25 presents the distribution of the standardised results in the written component for the general sample.



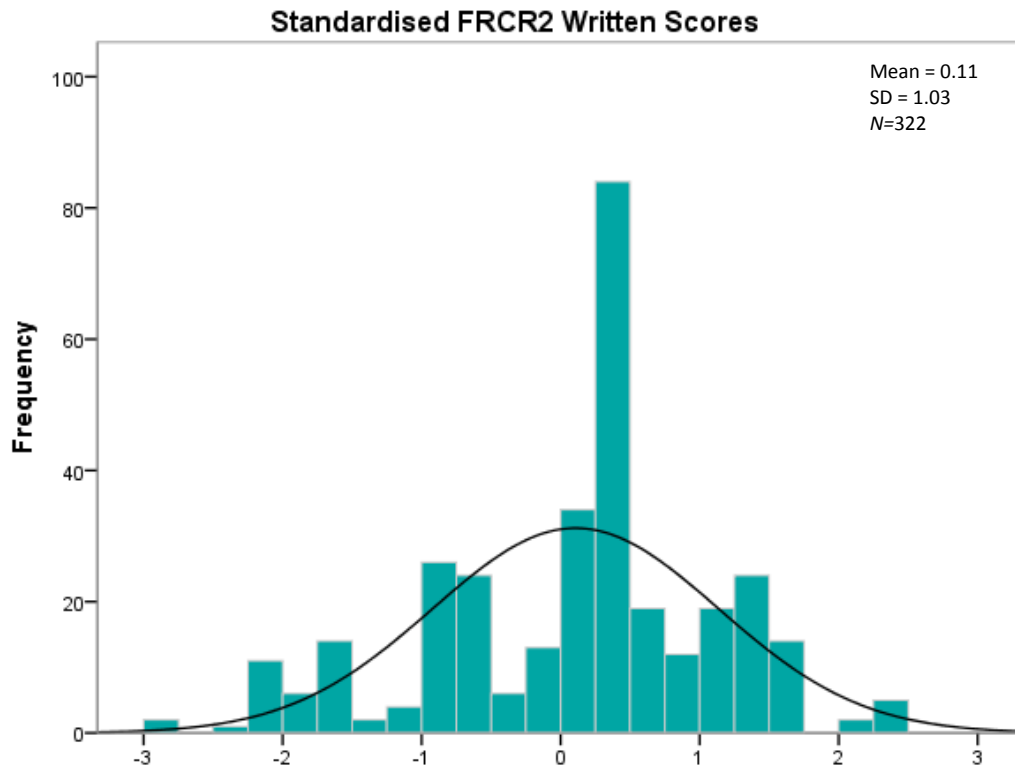


Figure 25. Distribution of FRCR2 written component standardised scores.

The distribution with a mean of 0.11 and a standard deviation of 1.03 indicates the results were skewed to the left ( $Z_{KS} = 2.93$ ,  $p < 0.001$ ). This translates into more candidates obtaining middle grades such as 'C' corresponding to approximately +0.28 SD across diets, which would explain the high peak around this value. Secondly, Z-transformation was performed on all candidates' attempts in FRCR2 (891 records), but the records of some of those candidates could not be matched with the History File, as their GMC Number was not available. Considering that these might have been the international candidates who were also more likely to obtain lower results (Hawtin *et al.*, 2014), this could have further skewed the results and increased the overall mean. Analogously, the distributions of the oral (and also clinical) components were found to be statistically different from normal.

### 5.3.2 Inferential statistics

The previous sections aimed to present a broad picture of the provided FRCR data. In light of the hypothesis set in the introductory part of this thesis, however, the purpose of this project was to establish whether there was a relationship between performance in the MRCP(UK) parts and the FRCR examination. This relationship was sought via correlation

coefficients, contrasting groups methods, and through fitting regression models, all of which constitute the next parts of this section.

### 5.3.2.1 Correlations

The relationships between MRCP(UK) scores and measures of performance in FRCR examinations were quantified using Pearson's product-moment correlation coefficient ( $r$ ). Table 42 presents the values of the coefficients for performance measures in FRCR1 and FRCR2 matched against MRCP(UK) scores. The analyses excluded FRCR1 censored candidates.

**Table 42. Correlation coefficients (Pearson's  $r$ ) between MRCP(UK) and performance measures in FRCR1 and FRCR2.**

<i>Variable</i>	<i>Part I Mark on 1<sup>st</sup> Attempt</i>	<i>Part II Mark on 1<sup>st</sup> Attempt</i>	<i>PACES Mark on 1<sup>st</sup> Attempt</i>
Mean Module Mark on 1 <sup>st</sup> Attempt in FRCR1	0.52** ( $n=177$ )	0.41** ( $n=203$ )	0.36** ( $n=209$ )
Total Number of Attempts in FRCR1	-0.36** ( $n=243$ )	-0.42** ( $n=320$ )	-0.27** ( $n=447$ )
Total Number of Attempts in FRCR2	-0.08 ( $n=58$ )	-0.37** ( $n=132$ )	-0.04* ( $n=245$ )
Pass/Fail in FRCR2	0.36** ( $n=58$ )	0.32** ( $n=133$ )	0.01 ( $n=245$ )
FRCR2 Written standardised score	0.39* ( $n=58$ )	0.53** ( $n=129$ )	0.03 ( $n=241$ )
FRCR2 Oral standardised score	0.20 ( $n=53$ )	0.33** ( $n=129$ )	0.12 ( $n=241$ )

\*\* significant  $p<0.001$ , \*  $p<0.05$

Inspection of Table 42 shows that the majority of the relationships were significant. The magnitude of the coefficients interpreted in accordance with Cohen's guidelines showed three coefficients being strong. These were the coefficients between Mean Module Mark on 1<sup>st</sup> Attempt in FRCR1 and Part I and Part II scores, and between Total Number of Attempts in FRCR1 and Part II scores.

The Mean Module Mark on 1<sup>st</sup> Attempt and the MRCP(UK) scores were continuous and allowed for more variability than, for example, Pass/Fail FRCR2 result or Total Number of Attempts in FRCR1 or FRCR2, which have resulted in relatively higher coefficients more accurately describing the relationships. Coefficients associated with these two FRCR performance measures were also positive, meaning that the higher the score in MRCP(UK), the higher the score in the FRCR1 and FRCR2 exams. Results with a negative sign were

associated with coefficients between the score measures and the Total Number of Attempts measures, which is intuitive; the higher the score, the fewer the attempts required to pass. Weak relationships (of approximately 0.10) were observed between the oral FRCR2 component and Part I and PACES scores; however, these relationships were statistically non-significant. Even smaller coefficients were observed between PACES and FRCR measures of performance, and between Part I and Total Number of Attempts in FRCR2. Eight of the twenty-one calculated coefficients could be considered moderate in magnitude.

On a more detailed level, the relationships between MRCP(UK) parts and separate module scores were tested. Pearson's *r* coefficients are provided in Table 43.

**Table 43. Correlation coefficients (Pearson's *r*) between MRCP(UK) scores and scores in FRCR1 modules**

<i>Variable max n=199</i>	<i>Part I Mark on Attempt1</i>	<i>Part II Mark on Attempt1</i>	<i>PACES Mark on Attempt1</i>
<b><u>Total Number of Attempts in</u></b>			
Cancer Biology	-0.27**	-0.37**	-0.20**
Clinical Pharmacology	-0.36**	-0.34**	-0.24**
Medical Statistics	-0.32**	-0.32**	-0.19**
Physics	-0.29**	-0.21**	-0.18**
<b><u>Module Mark on 1st Attempt in</u></b>			
Cancer Biology	0.41**	0.36**	0.27**
Clinical Pharmacology	0.47**	0.33**	0.30**
Medical Statistics	0.41**	0.32**	0.27**
Physics	0.41**	0.31**	0.30**

In accordance with Cohen's guidelines, the coefficients presented in Table 43 should be considered moderate in magnitude. The highest value coefficients were observed between marks in FRCR1 modules and Part I scores, but only slightly higher than the ones observed for Part II. The coefficients associated with PACES were even smaller.

#### ***Correction of correlation coefficients for range restriction and attenuation***

The correlation coefficients between MRCP(UK) parts and the FRCR tests were corrected for range restriction and disattenuated. Table 44 summarizes the standard deviations and the reliability coefficients required for the process. The unrestricted sample comprised all

candidates in the History File who had a score in an MRCP(UK) part. Each part was treated separately. It was not taken into account that Part II and PACES were first restricted by previous MRCP(UK) parts. The restricted sample standard deviations were calculated for each MRCP(UK) part only for those candidates who attempted FRCR. The table also contains the corrected correlation coefficients.

Correction was applied to continuous variables where the correlation coefficients were statistically significant and where the reliability coefficients were known. Additionally, for the purposes of meta-analyses, correction was also applied to the FRCR2 Pass/Fail outcome, and in the absence of published information on the reliability of this exam, it was assumed to be 0.80 (the reliability coefficient was assumed to be similar to the one of PACES and MRCGP CSA). Therefore, corrections were applied to the coefficients associated with the Mean Module Mark on 1<sup>st</sup> Attempt, FRCR2 written standardised score, and FRCR2 Pass/Fail outcome.

**Table 44. Derestriction of range and disattenuation of correlation coefficients between MRCP(UK) and selected FRCR performance measures with parameters required for corrections.**

<i>Dependent variable ('Y')</i>	<i>r</i>	<i>Average Reliab. of Y</i>	<i>SD restricted sample</i>	<i>SD unrestricted sample</i>	<i>Corrected coefficient</i>	<i>% change</i>
<b><u>Part I (mean reliability=0.91)</u></b>						
FRCR1 Mean Module Mark*	0.52	0.73	9.25	11.98	<b>0.74</b>	42%
FRCR2 written	0.39	0.85	6.91	11.98	<b>0.66</b>	69%
FRCR2 Pass/Fail	0.36	0.80	7.26	11.98	<b>0.62</b>	71%
<b><u>Part II (mean reliability=0.81)</u></b>						
FRCR1 Mean Module Mark*	0.41	0.73	6.45	7.62	<b>0.59</b>	47%
FRCR2 written	0.53	0.85	5.79	7.62	<b>0.75</b>	42%
FRCR2 Pass/Fail	0.32	0.80	5.67	7.62	<b>0.51</b>	59%
<b><u>PACES (mean reliability=0.82)</u></b>						
FRCR1 Mean Module Mark*	0.36	0.73	5.89	6.93	<b>0.52</b>	47%
FRCR2 written	n.s.	n/a	n/a	6.93	n/a	n/a
FRCR2 Pass/Fail	n.s.	n/a	n/a	6.93	n/a	n/a

\*Mean Module Mark on 1<sup>st</sup> Attempt in FRCR1

As in the case of previous analyses, the correction resulted in a significant increase in the correlation coefficients.

### 5.3.2.2 *Contrasting groups*

Further evidence of the relationship between MRCP(UK) performance and FRCR1 was obtained through a comparison of means between contrasting groups with FRCR1 Rank as the factor. The comparisons excluded the ‘censored’ candidates. The central tendency and dispersion measures for each FRCR1 Rank group are presented in Table 45.

**Table 45. Mean scores (with SDs) in the MRCP(UK) parts in division by three FRCR1 Rank groups**

<i>FRCR1 Rank</i>	<i>Part I 1<sup>st</sup> Attempt Mean and SD</i>		<i>Part II 1<sup>st</sup> Attempt Mean and SD</i>		<i>PACES 1<sup>st</sup> Attempt Mean and SD</i>	
Dropouts & Failed	-8.76	9.61	0.30	5.32	-1.39	6.75
Typical candidates	0.48	8.36	4.53	5.91	0.72	5.81
Highfliers	4.46	8.20	6.69	6.82	2.98	4.90

Based on the values from Table 45, the Dropouts & Failed group obtained the lowest results in all three parts of the MRCP(UK) examination, while the Highfliers obtained the highest. Three one-way ANOVA tests were performed on first attempt MRCP(UK) scores with FRCR1 Rank as the factor to confirm the statistical differences observed between these groups. The results indicated these differences were highly significant: for Part I  $F(2, 192)=15.99$ ,  $p<0.001$ , for Part II  $F(2, 269) = 7.20$ ,  $p=0.001$ , and for PACES  $F(2, 398)= 8.64$ ,  $p<0.001$ . The variances for the MRCP(UK) scores were assumed to be homogenous based on the values of the Levene’s test ( $F(2, 192)=0.17$ ,  $p=0.84$ ,  $F(2, 269)=1.68$ ,  $p=0.19$  and  $F(2, 398)=3.02$ ,  $p=0.05$ , respectively).

The post-hoc tests performed concurrently with ANOVA on the Part I and Part II scores indicated that the Dropouts & Failed group differed significantly from Typical candidates both in Part I ( $p<0.001$ ) and Part II results ( $p=0.013$ ). Similarly, Typical candidates differed significantly from Highfliers ( $p=0.004$  and  $p=0.013$ ), and the Dropouts & Failed group differed from the Highfliers ( $p<0.001$  in the case of both parts).

Analysis of PACES scores showed that the Dropouts & Failed group did not differ significantly from the Typical candidates ( $p=0.239$ ), but the Highfliers differed significantly from both Dropouts & Failed and Typical Candidate groups ( $p=0.003$  and  $p=0.001$ ).

Among the candidates for the FRCR, there were 83 MRCP(UK) Highfliers. Independent samples t-tests were employed to test the differences in FRCR performance between them and the rest of the candidates. The results indicated that MRCP(UK) Highfliers scored significantly higher in the Mean Module Mark on 1<sup>st</sup> attempt ( $t(229)=-5.01$ ,  $p<0.001$ ,  $r=0.31$ ), and in the FRCR2 written exam ( $t(320)=-2.10$ ,  $p=0.037$ ,  $r=0.11$ ). There was no statistically significant difference in the FRCR2 oral examination ( $t(319)=-1.62$ ,  $p=0.107$ ).

The FRCR1 Rank was also cross-tabulated with MRCP(UK) Highfliers to see if there was any congruence. Surprisingly, there were two MRCP(UK) Highfliers who failed FRCR1, 57 who were Typical FRCR1 candidates, and 27 who were also considered FRCR1 Highfliers. An independent samples t-test was performed on first attempt MRCP(UK) scores between those candidates who belonged to both Highfliers groups and those who belonged to just one. Indeed, the first group scored statistically significantly higher in all parts of MRCP(UK) (Part I:  $t(110)=-3.73$ ,  $p<0.001$ ,  $r=0.33$ , Part II:  $t(125)=-2.50$ ,  $p=0.014$ ,  $r=0.21$ , and PACES:  $t(164)=-3.54$ ,  $p=0.001$ ,  $r=0.27$ ). They were also significantly better at FRCR1 Mean Module Mark on 1<sup>st</sup> Attempt ( $t(109)=-2.90$ ,  $p=0.004$ ,  $r=0.27$ ) and in the written component of FRCR2 ( $t(115)=-2.36$ ,  $p=0.020$ ,  $r=0.21$ ), but no statistically significant difference was observed for the oral component of FRCR2 ( $p=0.116$ ).

#### **5.3.2.3 Regression models**

In order to establish the joint influence of MRCP(UK) parts on FRCR examination performance several regression models were fitted: three linear regression models and a logistic regression model. The models also excluded cases of the 'censored' candidates.

The first model regressed the first attempt scores in MRCP(UK) parts onto FRCR1 Mean Module Mark on 1<sup>st</sup> Attempt. The second and third models were used to estimate the effect of first attempt scores in MRCP(UK) parts onto standardised written and oral scores in FRCR2 assessments. In each case the entry method was used. Table 46 summarises the three models.

**Table 46. Summary of the regression models for selected FRCR performance measures as dependent variables .**

<i>Model</i>	<i>Dependent variable</i>	<i>Standardised Coefficients for</i>			<i>R<sup>2</sup></i>	<i>N</i>	<i>Average VIF</i>
		<i>Part I</i>	<i>Part II</i>	<i>PACES</i>			
model 1	Mean Module Mark on 1 <sup>st</sup> Attempt	0.31**	0.16	0.23*	0.295	159	1.46
model 2	Standardised score in Written FRCR2	0.09	0.43*	0.12	0.254	51	1.29
model 3	Standardised score in Oral FRCR2	-0.04	0.21	0.03	0.036	51	1.29

\*\* significant with  $p < 0.001$ , \* $p < 0.05$

The best model was fitted for FRCR1 Mean Module Mark. It was found that Part I and PACES were independent predictors of FRCR1 performance. The standardised beta coefficient for Part I was 0.31 ( $p < 0.001$ ), and for PACES it reached 0.23 ( $p = 0.001$ ). The model for the oral FRCR2 scores did not fit well; it explained little variance and the predictors were non-significant. Only the written component of the FRCR2 seemed to be predicted by the Part II results. The beta coefficient for Part II reached 0.43 and was highly significant ( $p = 0.006$ ). Based on the average VIF value, none of the models violated the assumption of multicollinearity.

The lack of significance of the coefficients in the third model might have been a result of the low number of valid cases for the analysis, due to the low number of candidates who had a record of Part I in the dataset, or alternatively, due to the oral scores having low reliability. The information on the reliability of the oral examination could not have been confirmed; however, it was possible to verify if the low number of cases affected the analyses. The EM algorithm was applied to estimate the missing values in MRCP(UK) scores (68.3% for Part I, 58.1% for Part II and 40.9 % for PACES) in order to fit the models again. The summary of the fitted models is presented in Table 47.

**Table 47. Summary of the regression models for selected FRCR performance measures with the EM algorithm imputed values.**

Model	Dependent variable	Standardised coefficients for			$R^2$	N	Avg. VIF
		Part I	Part II	PACES			
model 1*	Mean Module Mark on 1 <sup>st</sup> Attempt	0.30**	0.15	0.21**	0.277	231	1.65
model 2*	Standardised score in Written FRCR2	0.06	0.28**	-0.07	0.100	322	2.35
model 3*	Standardised score in Oral FRCR2	0.03	0.20*	0.04	0.059	321	2.35

\*\* significant with  $p < 0.001$ , \* with  $p < 0.05$

The models 1\* and 2\* with missing values imputed with EM algorithm fitted slightly less well to the data than the original models 1 and 2. However, the EM algorithm application allowed model 3\* to reveal the influence of Part II results on the oral FRCR2 examination.

The last model regressed first attempt scores in MRCP(UK) parts onto binary FRCR2 Pass/Fail result using the entry method. Based on the Hosmer-Lemeshow statistics it was well-fitted ( $\chi^2(8, n=53)=6.63, p=0.577$ ). Table 48 summarises the model (model 4).

**Table 48. Summary of the logistic regression models summary for FRCR2 Pass/Fail score before and after imputing missing values using EM algorithm.**

Model	Dependent variable	Predictors (odds ratios)			N
		Part I	Part II	PACES	
model 4	FRCR2 Pass/Fail score	1.06	1.20	1.05	53
		$p=0.44$	$p=0.07$	$p=0.53$	
		Wald : 0.60	3.23	0.39	
model 4*	FRCR2 Pass/Fail score	2.10	3.09*	0.77	337
		$p=0.34$	$p=0.04$	$p=0.48$	
		Wald : 0.89	4.36	1.20	

Model 4: Cox and Snell pseudo- $R^2 = 12\%$ , Model 4\*: pseudo- $R^2 = 6\%$ .

None of the predictors were significant; however, there was an almost significant effect of Part II results ( $p=0.07$ ), which would also be the best predictor in the model based on the value of the Wald statistic. The model was re-fitted using data estimated with the EM algorithm (model 4\* in the table above). Model 4\* also fitted well based on the Hosmer-Lemeshow test ( $\chi^2(7, n=337)=6.55, p=0.478$ ) and showed that Part II scores were a significant predictor of the FRCR2 ultimate Pass/Fail outcome. The odds ratio suggests that



with every standard deviation increase in Part II scores, the likelihood of passing FRCR2 increased nearly three times.

#### 5.3.2.4 Similarity of the FRCR models to the SCEs model

The differences between significant FRCR models (for Mean Module Mark on 1<sup>st</sup> Attempt and FRCR2 written exam scores) and the previously fitted SCEs general model were assessed using the Chow test. The general SCEs model comprised CKBA scores, as mentioned in the CKBA section above. The Chow test comparison was performed on Z-transformed scores, and the data required for the calculation are presented in Table 49.

**Table 49. Comparison between FRCR models and the SCEs & Cardiology joint model using Chow test.**

<i>Dependent variable</i>	<i>Residual Sum of Squares</i>	<i>N</i>	<i>No. of params</i>	<i>Value of F(df1. df2) Chow test</i>	<i>Critical Value</i>
<b><u>FRCR1 versus SCEs</u></b>					
FRCR1 Mean Module Mark on 1 <sup>st</sup> Attempt (model 1)	92.59	159	4	$F(4,1724) = 0.65$	2.34
SCEs general model	927.36	1573			
Joint model	1021.48	1732			
<b><u>FRCR2 versus SCEs</u></b>					
FRCR2 Written Score (model 2)	36.36	51	4	$F(4,1616) = 0.46$	2.38
SCEs general model	927.36	1573			
Joint model	964.81	1624			

Table 49 contains the results of the Chow tests ( $F$  statistic), which did not exceed the critical values. Therefore, there was no evidence that FRCR models differed significantly from the general SCEs model (with Cardiology).

## 5.4 MRCGP APPLIED KNOWLEDGE TEST

The Membership of the Royal College of General Practitioners consists of two assessments: Applied Knowledge Test ('AKT') and Clinical Skills Assessment ('CSA'). AKT is a written knowledge test comprising three groups of questions: evidence investigation, clinical medicine and organisational questions, with 80% of questions referring to clinical medicine. Its purpose is to assess higher order reasoning and problem solving. As such, AKT was considered to be related to the knowledge component of professionalism rather than clinical skills. It resembles MRCP(UK) Part II more than it does Part I. More on the exam can

be found in section 2.4.2. CSA examination is a clinical skills assessment and while descriptive statistics are presented here, the inferential analyses are provided separately in Chapter 6.

#### 5.4.1 Descriptive statistics

The analysed dataset contained 2,284 records of the first attempt results in the CSA and AKT. The CSA descriptive statistics are provided here to introduce the MRCGP exam, and to avoid repetitiveness of information in Chapter 6, where the inferential statistics for CSA are discussed.

In the first step, the candidates were divided based on the order in which they had attempted the exams, i.e. into those who attempted MRCP(UK) first and those who attempted MRCGP first. For inferential analyses, those candidates who attempted MRCGP first had to be excluded in accordance with the aims of this study.

Table 50 summarizes the numbers of candidates who passed or failed the MRCP(UK) and MRCGP. Failure in the case of MRCGP was defined as failing the first attempt in AKT or CSA, as the dataset provided only first attempt results. In the case of MRCP(UK), a failure was defined as not having passed all parts of MRCP(UK) in the defined time-frame.

**Table 50. Cross-table for MRCP(UK) and MRCGP Pass/Fail outcomes.**

	<b>MRCP(UK) failed</b>		<b>MRCP(UK) passed</b>		<b>TOTAL:</b>	
	<i>MRCP(UK) taken first</i>	<i>whole sample</i>	<i>MRCP(UK) taken first</i>	<i>whole sample</i>	<i>MRCP(UK) taken first</i>	<i>whole sample</i>
<b>MRCPGP failed</b>	1,256	1,400	604	739	1,860	2,139
<b>MRCPGP passed</b>	111	135	5	10	116	145
<b>TOTAL:</b>	1,367	1,535	609	749	1,976	2,284

Inspection of Table 50 shows that the majority of the candidates in the dataset were still in the process of passing either of the exams. There was no significant difference in the likelihood of passing AKT between doctors who attempted MRCP(UK) first or MRCGP first (odds ratio 1.11;  $\chi^2(1, n=2,284)=0.35, p=0.555$ ).

The final sample consisted of 1,976 cases, after excluding doctors ( $n=308$ ) who attempted MRCP(UK) after attempting MRCGP. Of those, 44.4% (878) were male candidates and 55.6% (1,098) were female. Being other than white ethnicity was declared by 60.4% (1,194)

candidates. The majority of them obtained their primary qualification in the UK (63.8%; 1,261) and nearly 99.5% of them (1,966) had both a GMC Number and a UK address (UK trainee). There were no missing MRCP(UK) Part I scores in the dataset; however, censoring affected Part II and PACES first attempt scores. The dataset contained 938 candidates with a valid Part II scores and 739 candidates with PACES scores.

The distributions of the attempt results in Part I and PACES were significantly different from normal in the analysed sample, as indicated by the value and significance of the K-S tests ( $Z_{KS} = 1.39$ ,  $p=0.042$ ;  $Z_{KS} = 1.97$ ,  $p=0.001$ ). The distribution of Part II scores was close to normal ( $Z_{KS} = 0.83$ ,  $p=0.492$ ). The distributions of the AKT scores and its subtests were also significantly different from normal ( $p<0.001$ ). Despite these results, further analyses were performed using parametric tests, for the same reasons as previously argued for (see Chapter 4 for details).

## 5.4.2 Inferential statistics

### 5.4.2.1 Correlations

The correlations coefficients between MRCP(UK) first attempt scores and AKT first attempt scores (including its parts) were calculated and are presented in Table 51.

**Table 51. Pearson's product moment correlation coefficients between MRCP(UK) parts scores and AKT parts scores.**

<i>MRCP(UK) Part</i>	<i>AKT Raw Mark</i>	<i>AKT Clinical Medicine Mark (%. not scaled)</i>	<i>AKT Evidence Interpretation ('research') Mark (%. not scaled)</i>	<i>AKT Organisational Questions Mark (%. not scaled)</i>
Part I	0.66** ( <i>n</i> =1,976)	0.66** ( <i>n</i> =1,976)	0.46** ( <i>n</i> =1,976)	0.31** ( <i>n</i> =1,976)
Part II	0.59** ( <i>n</i> =938)	0.58** ( <i>n</i> =938)	0.41** ( <i>n</i> =938)	0.32** ( <i>n</i> =938)
PACES	0.43** ( <i>n</i> =739)	0.41** ( <i>n</i> =739)	0.33** ( <i>n</i> =739)	0.22** ( <i>n</i> =739)

\*\* significant with  $p<0.001$

Inspection of Table 51 shows that the higher the scores in the MRCP(UK) parts, the higher the AKT scores. Part I and Part II correlated stronger with the overall AKT Raw Mark ( $r=0.66$  and  $r=0.59$ , respectively) than PACES ( $r=0.43$ ), and the difference between these correlation coefficients was statistically significant as tested with Fisher's  $r$ -to- $z$  transformation.

Further inspection revealed that the coefficients between AKT Clinical Medicine ('AKT CM') and MRCP(UK) parts were almost equal to the ones for the overall AKT Raw Mark. All differences in the values of those coefficients between MRCP(UK) parts were statistically significant as well. AKT Evidence Interpretation ('AKT EI') also correlated highest with Part I; however, there was no statistical difference between the coefficient and the one obtained for Part II results ( $Z=1.55$ ,  $p=0.12$ ). The relationship between AKT EI and PACES was significantly weaker than between AKT EI and Part I and Part II.

The relationships between AKT Organisational Questions ('AKT OQ') and MRCP(UK) parts were the smallest in magnitude; the correlation coefficients were much smaller than in the case of AKT CM and AKT EI. The coefficients between AKT OQ and the three parts of MRCP(UK) were similar in size.

### ***Range restriction correction and disattenuation of the correlation coefficients***

Although the above-given correlation coefficients did not require correction for restriction of range, as neither of the exams limits the right to attempt it based on the results of the other, the correction was introduced due to the self-selection process that takes place in the case of these two exams. The coefficients were also disattenuated. Table 52 contains all values of parameters required for the range restriction correction and disattenuation. The unrestricted sample in this case comprised all candidates having records in the History File for each part of the exam, while the restricted sample comprised only those candidates that have also attempted MRCGP.

**Table 52. Range derestriction and disattenuation of the correlation coefficients between MRCP(UK) and AKT, with parameters required for both corrections.**

<i>MRCP(UK) Part</i>	<i>r</i>	<i>Average Reliab. of MRCP part</i>	<i>Average Reliab. of AKT</i>	<i>SD restricted sample</i>	<i>SD unrest. sample</i>	<i>Corrected coefficient (r)</i>	<i>% change</i>
Part I	0.66	0.91	0.89	10.36	11.98	<b>0.79</b>	19%
Part II	0.59	0.81	0.89	6.58	7.62	<b>0.76</b>	28%
PACES	0.43	0.82	0.89	6.59	6.93	<b>0.53</b>	23%

Corrections resulted in an increase in the value of the original coefficients by 19 to 28% with the relationships becoming very strong.

### 5.4.2.2 Contrasting groups

In the MRCGP dataset, a total of 263 MRCP(UK) Highfliers was identified. Independent samples t-tests were used for comparison of mean AKT scores between the group of MRCP(UK) Highfliers and the rest of the candidates. These differences were statistically significant, and are summarised in Table 53. The table also provides the means and SDs, numbers of valid cases in both groups, and the estimated effect sizes.

**Table 53. Comparison of mean AKT scores between the MRCP(UK) Highfliers and Typical Candidates.**

AKT score	Means and SDs		Independent samples t-test	N Highfliers versus N Rest	Effect size (r)
	Rest of the sample	Highfliers			
AKT Raw Mark	152.00 (SD=14.70)	169.27 (SD=9.66)	$t(1,974) = -18.44,$ $p < 0.001$	263/1,713	0.38
AKT CM	77.80 (SD=7.24)	86.21 (SD=4.82)	$t(1,974) = -18.24,$ $p < 0.001$	263/1,713	0.37
AKT EI	73.49 (SD=15.11)	85.69 (SD=10.37)	$t(1,974) = -12.64,$ $p < 0.001$	263/1,713	0.27
AKT OQ	68.42 (SD=12.78)	75.86 (SD=11.32)	$t(1,974) = -8.91,$ $p < 0.001$	263/1,713	0.19

The results of the comparisons have shown that MRCP(UK) Highfliers scored significantly better in all AKT assessments.

### 5.4.2.3 Regression models

Several linear regression models regressed MRCP(UK) parts scores onto AKT scores and its partial scores. Separate models were fitted for: AKT Raw Mark, AKT Clinical Medicine, AKT Evidence Interpretation, and AKT Organisational Questions using the entry method. Table 54 provides the summary of these models.

**Table 54. Summary of the linear regression models for MRCGP AKT scores.**

<i>Model</i>	<i>Dependent variable</i>	<i>Standardised coefficients for</i>			<i>R<sup>2</sup></i>	<i>N</i>	<i>Average VIF</i>
		<i>Part I</i>	<i>Part II</i>	<i>PACES</i>			
model 1	AKT Raw Mark	0.32**	0.33**	0.19**	0.460	738	1.39
model 2	AKT CM	0.32**	0.32**	0.18**	0.423	738	1.39
model 3	AKT EI	0.21**	0.22**	0.18**	0.229	738	1.39
model 4	AKT OQ	0.14**	0.22**	0.09*	0.132	738	1.39

\* statistically significant with  $p<0.05$ , \*\*statistically significant with  $p<0.001$

The best model assumed AKT Raw Mark as the dependent variable; it explained 46% of variance. In this model, Part I and Part II scores were both stronger predictors than PACES scores (beta=0.32 and beta=0.33 *versus* beta=0.19, respectively). Models fitted for partial AKT assessments varied between one another. The best model explained 42.3% of variance and was fitted for AKT CM scores. This model was nearly identical to the AKT Raw Mark model. The models with AKT EI scores and AKT OQ scores as dependent variables were less well fitted (22.9% and 13.2% of variance explained, respectively). In the AKT EI model, all MRCGP(UK) parts had an almost equal effect based on the values of beta coefficients (0.21, 0.22 and 0.18, respectively), while in the case of the AKT OQ model the strongest predictor was Part II (beta=0.22). In this last model, Part I and PACES scores were substantially weaker predictors (beta=0.14 and beta=0.09, respectively).

#### **5.4.2.4 Similarity of the linear regression models**

The AKT Raw Mark model was tested for similarity to the SCEs (with Cardiology) model and the FRCR1 Mean Module Mark on first Attempt model (as the best fitted model among FRCR models). The comparison was performed on the standardised AKT scores using the Chow test. Table 55 presents the residuals required for calculation of the *F* statistic, valid numbers of cases, and the summary of the Chow tests.

**Table 55. Comparison of AKT, SCEs and FRCR1 models using Chow test.**

<i>Dependent variable</i>	<i>Residual Sum of Squares</i>	<i>N</i>	<i>No. of params</i>	<i>Value of F(df1, df2) Chow test</i>	<i>Critical Value</i>
<b><u>AKT versus SCEs</u></b>					
AKT Raw Mark model	501.37	1,476	4	<b><i>F(4, 3,041)=92.75</i></b>	2.37
SCEs with Cardiology	927.36	1,573			
Joint model	1,603.03	3,049			
<b><u>AKT versus FRCR1</u></b>					
AKT Raw Mark model	501.37	1,476	4	<b><i>F(4, 1,627) = 322.08</i></b>	2.38
FRCR1 model	92.59	159			
Joint model	1.064.28	1,635			

The results of the Chow tests suggest that AKT model was statistically significantly different from the previously fitted models for both FRCR1 and SCEs (including Cardiology).

## **SUMMARY AND DISCUSSION**

The hypothesis was that MRCP(UK) would predict performance in knowledge exams attempted by doctors a few years after taking MRCP(UK). The evidence collected in this chapter supported this hypothesis, as MRCP(UK) parts scores were all positively associated with actual scores in all knowledge exams that were investigated. This relationship was expressed in the values of the coefficients and their statistical significance, both in the univariate and multivariate analyses.

The univariate statistics have shown that there was a positive correlation between MRCP(UK) parts scores and scores achieved in other exams, meaning that the better the candidate performance in MRCP(UK), the better their score in the other exams. The strength of these relationships was rather moderate and varied depending on the MRCP(UK) part and subsequent measure. The correlation coefficients associated with Part I and Part II were generally higher than those associated with PACES. The uncorrected coefficients ranged from 0.29 to 0.66 for Part I, from 0.42 to 0.66 for Part II, and from 0.16 to 0.50 for PACES. Corrections for attenuation and range restriction further increased these values substantially, making them in some cases extremely high. However, the corrections merely aim to approximate the strength of the true relationship between the predictor and the criteria, and the magnitude of the corrected coefficients should not be interpreted as

definite. Corrections were necessary from the perspective of meta-analysis as discussed in section 3.7.9 of the Methodology chapter. However, the interpretation of the results seems to be more intuitive when observed (uncorrected) correlation coefficients are discussed.

The comparison of means provided further evidence in favour of the validity of MRCP(UK). Firstly, the FRCR1 Highfliers scored significantly better in all MRCP(UK) parts than Typical candidates and those who Failed or Dropped out. Secondly, the group of MRCP(UK) Highfliers scored significantly better in all subsequent knowledge exams: SCEs, Cardiology, FRCR and MRCGP AKT. It may be that doctors belonging to the group of Highfliers are simply more talented or more intelligent, and that the observed relationships have more to do with their high aptitude rather than the properties of the tests. However, the example of the observed relationship between MRCP(UK) Highfliers and FRCR1 Rank Highfliers indicates that higher aptitude is not an explanation. It was found that among the MRCP(UK) Highfliers, there were several candidates considered Typical in terms of FRCR1 examinations, and there were even two cases of failed FRCR1 candidates. A remarkable success in one exam did not, therefore, automatically guarantee a remarkable success in another. This particular result stands in line with the previous research, which showed that aptitude does not predict future performance as well in comparison to educational attainment tests (McManus, Dewberry, *et al.*, 2013). This is also in line with the concept of the Academic Backbone referred to previously (McManus *et al.*, 2013), where the authors argue that aptitude is not the main foundation for success, but an aid in the process of accumulating knowledge. The variability in the test results among highfliers suggests that the exams indeed test for knowledge. Further evidence against the aptitude hypothesis comes from the comparison of the coefficients associated with AKT partial scores. Significant differences in the strength of the coefficients between AKT Organisational Questions, AKT Evidence Interpretation, and AKT Clinical Medicine were observed, with the latter being significantly higher. Also, the relationship between AKT Organisational Questions and Part I scores was weaker than that with Part II scores. Both observations suggest that the contents of the exam had a significant impact on the strength of the correlation coefficients; this would not be observed if aptitude was solely responsible for these relationships. Hence, the significant differences in performance between the Highfliers and Typical candidates support the exams validity.

The multivariate regression models fitted to assess the relative importance of MRCP(UK) parts on dependent variables further indicated that Part I and Part II scores had significantly higher impacts on measures of knowledge than PACES. This was supportive of the findings



of univariate analyses. Part II scores were a significant predictor in almost all regression models with exceptions of the Medical Oncology, FRCR1 examination, and Palliative Medicine (barely non-significant) models. They were also the sole significant predictor in the case of four of the models (Infectious Diseases, Renal Medicine, and FRCR2 written examination), and after implementation of the EM algorithm, they became the sole predictor for the FRCR2 oral examination. Whenever Part I and Part II scores were included as predictive of the dependent variable, the beta coefficients associated with the Part II scores were usually higher than those associated with the Part I scores. The exceptions to that rule were Gastroenterology, Cardiology, and FRCR1 models, where Part I coefficients were higher; and the MRCGP AKT, Dermatology, Neurology, and Respiratory Medicine models, where both Part I and Part II scores had a similar impact on the dependent variables. Finally, PACES scores were excluded from the majority of the models and whenever they were a significant predictor, which was the case for Acute Medicine, Endocrinology, Gastroenterology, Geriatric Medicine, Rheumatology, FRCR1 and Cardiology, they were always associated with a substantially smaller coefficient than Part I and Part II. The above results confirm a theoretical pattern of associations that was hypothesised based on the theoretical notions of psychometrics. Firstly, correlation between two similar constructs should be higher than between two dissimilar ones (Anastasi & Urbina, 1997) and Part I, Part II and all of the criterial exams are written knowledge exams, while PACES is a structured clinical assessment. Secondly, the similarity of the exams forms would also lead to higher coefficients (Cronbach, 1970), and Part I, Part II and the criterion measures were mostly written exams – with the exception of the FRCR2 oral exam – and aimed to assess knowledge in a written form. The predominant role of Part II and Part I scores over PACES in the prediction models for knowledge assessments can, therefore, find support in both theoretical notions. These arguments would find more support if a reversed pattern is observed between PACES and the clinical performance measures, which constitutes the subject of Chapter 6.

Despite the observed differences between the linear regression models, they were shown to be statistically similar, with the exception of the AKT model. This would indicate that MRCP(UK) scores predict subsequent exams scores to a similar extent. In other words, if the fitted multivariate regression models could have been drawn in a multi-dimensional space, their representations (lines, surfaces, etc.) would be parallel, as observed with the lines in the simplified Figure 20, and the multidimensional slopes would correspond to the magnitude of the relationships (Nathans, Oswald, & Nimon, 2012).

In conclusion, the results of analyses provided in this chapter support the predictive validity of the MRCP(UK) examination in terms of predicting the knowledge component of competence. The results suggest that MRCP(UK) Part I and Part II test for knowledge. Further, it can be inferred based on the results that criterial exams were mostly based on data interpretation (higher coefficients for Part II) rather than factual knowledge (Part I). Significant coefficients for PACES further suggest that the specialty exams and equivalent are to a certain extent based on clinical skills, or that PACES does require implementation of knowledge. The similarities between the linear models have shown that MRCP(UK) predicts subsequent knowledge tests to a certain extent only and that this effect could potentially be averaged out. The issue of estimating the average size of the relationship between MRCP(UK) parts and the criterion measures is the subject of Chapter 7 of this thesis, where meta-analytical models are fitted to the data.

## Chapter 6

### Assessment of Clinical Skills and Attitudes

#### ABSTRACT

*The hypothesis for this research did not solely address the issue of predicting knowledge measures, but also assumed that performance in MRCP(UK) would predict measures of clinical performance. Five such measures were identified in the course of this research: FRCR2 clinical exam performance, MRCGP CSA performance, two registration statuses based on LRMP (Licence Issues and Voluntary Erasures), being under review by the GMC FtP panel, and performance in higher specialty training assessed with Annual Reviews of Competence Progression (ARCP). The LRMP records supplied one additional measure, that of professional attitudes, which was being erased from LRMP for administrative reasons. This measure was regarded a distant proxy for conscientiousness. The univariate and multivariate analyses indicated that MRCP(UK) scores predicted the above-mentioned measures to a different extent, but still supporting the hypothesis. The results of this chapter are complementary to those of Chapter 5.*

As set out in Chapter 2, the predictive validity of MRCP(UK) was to be established through investigating two groups of evidence coherent with knowledge, and clinical skills and clinical performance including professional attitudes. The previous chapter (Chapter 5) focused solely on analysing the relationships between MRCP(UK) parts and criteria that represented knowledge. The purpose of this chapter is to complement these analyses by investigating the relationships between MRCP(UK) and measures of clinical performance and attitudes. Altogether six measures were identified: the FRCR2 clinical component, the MRCGP CSA examination, the registration statuses based on LRMP indicating either Issues with Licence, Administrative Erasures, or Voluntary Erasures, and also GMC FtP investigations. The analyses are presented in the following sections of this chapter.

#### 6.1 FRCR2 CLINICAL EXAMINATION

The FRCR2 exam apart from its oral and written components, also comprises a clinical component, one that similarly to PACES consists of five stations where a candidate needs to approach clinical scenarios. However, the scenarios differ significantly from those used in PACES. Although the ability to physically examine a patient is required of the candidate

clinical oncologists, they are also required to, for example, plan radiotherapy for a patient with a malignant tumour. Hence, the actual overlap of tested skills between the exams may be small. At the initial stages of this research during consideration of criteria, the FRCR2 clinical component was assumed to be a measure of clinical skills.

The FRCR2 sample size ( $n=218$ ) and its demographic characteristic were already provided in Chapter 5, section 5.3 on FRCR examination. The distribution of the clinical component scores was significantly different from normal, as found by one-sample K-S test ( $Z_{KS}= 3.13$ ,  $p<0.001$ ). This was similar for MRCP(UK) PACES scores ( $Z_{KS}=1.38$ ,  $p=0.045$ ). The values of the K-S tests indicated that Part I and Part II first attempt results were normally distributed ( $Z_{KS}=0.56$ ,  $p=0.916$  and  $Z_{KS}=0.63$ ,  $p=0.818$ , respectively). These results indicated the need for the use of non-parametric statistical tests; however, in view of the findings of Chapter 4, parametric methods were applied.

The FRCR2 clinical component analyses followed the same logical order as for the other components. It was found that the FRCR1 Highfliers ( $n=102$ ,  $M=0.38$ ,  $SD=0.89$ ) scored significantly higher in FRCR2 clinical exam than Typical candidates ( $n=218$ ,  $M=0.05$ ,  $SD=0.97$ ), as confirmed by independent samples t-test ( $t(318)=-2.97$ ,  $p=0.003$ ,  $r=0.16$ ).

Further it was found that the FRCR2 clinical component correlated with the first attempt Part II scores ( $r=0.21$ ,  $p=0.019$ ), but the magnitude of this relationship was rather small. Correlation between FRCR2 clinical exam and Part I reached 0.20 ( $p=0.149$ ) and with PACES was only 0.08 ( $p=0.206$ ). In order to assess the joint effect of MRCP(UK) parts on FRCR2 clinical scores a linear regression model fitting was attempted using the entry method. None of the predictors were significant. Application of the EM algorithm to generate estimates for the missing values in MRCP(UK) parts scores to increase the sample size changed the relationships slightly. Table 56 contains the summary of the fitted regression models, where model 1 is the one based on 51 cases, while model 1\* is the model with missing values imputed with the EM algorithm.

The model 1\* explained barely any variance; however, the estimation of missing values shifted the importance of PACES scores towards Part II scores. The relationships were still non-significant. The tests for similarity of the FRCR2 clinical component model and the SCEs and MRCGP AKT models was not performed due to weak parameters.

**Table 56. Summary of the FRCR2 clinical examination linear regression models before and after EM algorithm.**

<i>Model</i>	<i>Dependent variable</i>	<i>Standardised coefficients for</i>			<i>R<sup>2</sup></i>	<i>N</i>	<i>Average VIF</i>
		<i>Part I</i>	<i>Part II</i>	<i>PACES</i>			
model 1	Standardised score in Clinical FRCR2	0.03	0.09	0.13	0.032	51	1.30
model 1*	Standardised score in Clinical FRCR2	0.04	0.11	0.03	0.020	321	2.35

## 6.2 MRCGP CLINICAL SKILLS ASSESSMENT

The Clinical Skills Assessment, also referred to as CSA, is a part of the examination for the Membership of the Royal College of General Practitioners, and in its form resembles PACES. It is an OSCE type examination with thirteen clinical scenarios, during which a candidate is required to make evidence-based clinical decisions based on their knowledge, while at the same time communicating effectively with patients played by actors (see section 2.4.2 for more details). CSA was assumed to be a measure of clinical performance, although the relationships with knowledge tests were also expected to be revealed.

### 6.2.1 Descriptive statistics

The description of the dataset was already provided in Chapter 5, section 5.4 on MRCGP AKT. The sample sizes for Clinical Structured Assessment were analogous to the ones obtained for AKT, and were also reduced by the order in which the MRCP(UK) and MRCGP were taken. The main analyses were performed on 1,976 cases, as 308 candidates who attempted MRCGP before MRCP(UK) were excluded. Of those who were included in the final sample, 509 candidates attempted CSA in the new scheme, while 1,467 attempted the CSA exam in its old format. All candidates had a record of the equated score. Records of the first attempt scores in Part I were available for all candidates in the dataset; however, due to data censoring only 938 candidates had a record of first attempt results in Part II, and only 739 candidates had such a record for PACES.

It is worth noting that those who attempted MRCP(UK) first were 2.29 times more likely (odds ratio) to pass the CSA (statistically significant,  $\chi^2(1, n=2284) = 42.71, p<0.001$ ) than those who attempted MRCGP first. This ratio is significantly higher than that obtained for AKT. The distributions of the CSA results in the old scheme, new scheme, and equated

scores were significantly different from normal based on the results of the K-S tests (new scheme:  $Z_{KS} = 1.51$ ,  $p=0.021$ , old scheme:  $Z_{KS} = 6.77$ ,  $p<0.001$ , equated scores:  $Z_{KS} = 5.11$ ,  $p<0.001$ ).

The distribution of Part II scores was close to normal ( $Z_{KS} = 0.83$ ,  $p=0.492$ ), while Part I and PACES scores distributions were significantly different from normal as tested with K-S test:  $Z_{KS} = 1.39$ ,  $p=0.042$  and  $Z_{KS} = 1.97$ ,  $p=0.001$ , respectively. Despite this, all analyses were performed using parametric tests for the reasons mentioned previously in Chapter 4.

## 6.2.2 Inferential statistics

### 6.2.2.1 Correlations

The correlation coefficients between MRCP(UK) scores and three CSA scoring types were calculated as Pearson's  $r$ , and are presented in Table 57.

**Table 57. Correlation coefficients (Pearson's  $r$ ) between MRCGP CSA and MRCP(UK) parts.**

<i>MRCP(UK) scores</i>	<i>CSA old scheme</i>	<i>CSA new scheme</i>	<i>CSA equated score</i>
Part I 1 <sup>st</sup> attempt	0.32** ( $n=1,467$ )	0.43** ( $n=509$ )	0.35** ( $n=1,976$ )
Part II 1 <sup>st</sup> attempt	0.36** ( $n=691$ )	0.46** ( $n=247$ )	0.38** ( $n=938$ )
PACES 1 <sup>st</sup> attempt	0.43** ( $n=553$ )	0.58** ( $n=186$ )	0.46** ( $n=739$ )

\*\* significant at  $p<0.001$

All coefficients showed a positive association between MRCP(UK) parts scores and CSA, suggesting that the higher the scores in MRCP(UK), the better the scores in CSA. The coefficients for the new scheme were higher than those observed for the old format CSA and the equated score. This stems directly from the new CSA being based on a better scale (ordinal rather than dichotomous), and there being a wider range of total scores for the new CSA, which increases the variance.

The coefficients for CSA and PACES were higher than corresponding coefficients for Part I and Part II. The observed differences in magnitude between coefficients were in general statistically non-significant with one exception: the coefficient for the equated CSA score and PACES differed significantly from the one obtained for Part I and Part II and CSA equated scores ( $p<0.05$ ). The highest values of the coefficients were observed for the new scheme CSA scores, probably due the fact it is a more varied measure.

### ***Correction of correlation coefficients for range restriction and attenuation***

The presented correlation coefficients did not require correction for the restriction of range as described previously in the section devoted to AKT, but nonetheless the correction was introduced due to auto-selection process of the candidates. Also correction for unreliability of the measures was employed. The summary of data required for derestriction and disattenuation, and the results of the correction are presented in Table 58. In the absence of published reliability coefficient for CSA at the time when these analyses were performed, the coefficient was assumed to reach 0.80.

**Table 58. Derestriction of range and disattenuation of the correlation coefficients between MRCP(UK) and CSA and the parameters required for both corrections.**

<i>Dependent variable</i>	<i>r</i>	<i>SD restricted sample</i>	<i>SD unrest. sample</i>	<i>Corrected coefficient</i>	<i>% change</i>
<b><u>Part I (mean reliability =0.91)</u></b>					
CSA old scheme	0.32	10.33	11.98	<b>0.43</b>	33%
CSA new scheme	0.43	10.44	11.98	<b>0.56</b>	30%
CSA equated score	0.35	10.36	11.98	<b>0.46</b>	32%
<b><u>Part II (mean reliability =0.81)</u></b>					
CSA old scheme	0.36	6.36	7.62	<b>0.52</b>	44%
CSA new scheme	0.46	7.11	7.62	<b>0.60</b>	31%
CSA equated score	0.38	6.59	7.62	<b>0.53</b>	39%
<b><u>PACES (mean reliability =0.82)</u></b>					
CSA old scheme	0.43	5.45	6.93	<b>0.64</b>	48%
CSA new scheme	0.58	6.87	6.93	<b>0.72</b>	25%
CSA equated score	0.46	6.59	6.93	<b>0.59</b>	29%

The corrections resulted in coefficients increasing on average by 34%.

#### ***6.2.2.2 Contrasting Groups***

In order to test if there were any significant differences between the MRCP(UK) Highfliers and Typical Candidates in their CSA scores, a comparison of means was performed using independent samples t-tests. There were 53 MRCP(UK) Highfliers attempting CSA in the new scheme, 210 in the old scheme, and 263 for the equated score (in total). The mean scores, results of the t-tests and the effect sizes are presented in Table 59.

**Table 59. Comparison of mean scores in CSA between MRCP(UK) Highfliers and Typical candidates.**

<i>Measure</i>	<i>Means and SDs</i>		<i>Independent samples t-test</i>	<i>Effect size (r)</i>
	<i>Highfliers</i>	<i>Typical candidates</i>		
CSA new scheme	17.36 (SD=10.06)	5.59 (SD=13.41)	$t(507)=-6.19, p<0.001$	0.54
CSA old scheme	10.70 (SD=1.44)	9.24 (SD=2.17)	$t(1,465)=-9.46, p<0.001$	0.49
CSA equated	13.85 (SD=9.78)	4.43 (SD=12.75)	$t(1,974)=-11.48, p<0.001$	0.52

The above-presented differences in the CSA scores were statistically significant irrespective of the assessment scheme, and the Highfliers always performed better than the Typical candidates.

### 6.2.2.3 Regression models

In order to estimate the joint effect of MRCP(UK) parts on the CSA scores, the MRCP(UK) first attempt scores were regressed onto CSA scores (in three variants) using entry method. The models are summarised in Table 60 (below).

**Table 60. Linear regression models for CSA scores with MRCP(UK) parts as predictors.**

<i>Model</i>	<i>Standardised coefficients for</i>			<i>N</i>	<i>R<sup>2</sup></i>	<i>Average VIF</i>
	<i>Part I</i>	<i>Part II</i>	<i>PACES</i>			
CSA new scheme	0.18*	0.18*	0.46**	185	43.1%	1.41
CSA old scheme	0.04	0.15**	0.36**	553	20.7%	1.37
CSA equated	0.08*	0.16**	0.38**	738	25.4%	1.39

\* significant with  $p<0.05$ , \*\* significant with  $p<0.001$ ,

Irrespective of the assumed scoring scheme, the best predictors were always PACES first attempt scores, with beta coefficients reaching from 0.36 to 0.46. In each case it was more than double the values of beta coefficients for Part I or Part II. The models met the assumption of a lack of multicollinearity based on the Average VIF values.

### 6.2.2.4 Similarity of the linear regression models

Three CSA models were compared using the Chow test. Table 61 summarises the residual sums of squares, counts required for the Chow comparison, and the Chow test results.



**Table 61. Results of the Chow test comparisons between models for the three CSA scoring schemes.**

<i>Dependent variable</i>	<i>Residual Sum of Squares</i>	<i>N</i>	<i>No. of parameters</i>	<i>Value of F(df1, df2) Chow test</i>	<i>Critical Value</i>
<b><u>New versus Old</u></b>					
CSA new scheme	102.45	185	4	<b>F(4,730)=2.66</b>	2.38
CSA old scheme	359.03	553			
Joint model	468.21	738			
<b><u>New versus Equated</u></b>					
CSA new scheme	102.45	185	4	F(4,915)=1.34	2.38
CSA equated score	503.22	738			
Joint model	609.21	923			
<b><u>Old versus Equated</u></b>					
CSA old scheme	359.03	553	4	F(4,1283)=0.50	2.37
CSA equated score	503.22	738			
Joint model	863.60	1291			

Inspection of Table 61 above shows that despite the models of the new and old scoring schemes were different from one another, none of them were significantly different from the CSA Equated scores model, supporting the applied equating procedure. The CSA Equated model was used for further comparisons with the models for the other exams.

The CSA Equated score model was tested against the AKT Raw Mark model, the SCEs general model, and the FRCR1 model. Cardiology results were included in the general SCEs model. The Chow test comparisons were performed on Z-transformed scores, as each of the exams were reported on a different scale. Table 62 presents the counts and residual sums of squares required for performing the Chow test, together with final results.

**Table 62. Comparisons between CSA Equated regression model and AKT, SCEs with Cardiology and FRCR1 models using the Chow test.**

<i>Dependent variable</i>	<i>Residual Sum of Squares</i>	<i>N</i>	<i>No. of parameters</i>	<i>Value of F(df1, df2) Chow test</i>	<i>Critical Value</i>
<b><u>AKT versus CSA</u></b>					
AKT Raw mark	501.37	1,476	4	<b><i>F(4,2206)=37.86</i></b>	2.38
CSA Equated score	503.22	738			
Joint model	1073.56	2,214			
<b><u>CSA versus SCEs</u></b>					
CSA Equated score	102.45	185	4	<b><i>F(4,2303)=22.42</i></b>	2.38
SCEs with Cardiology	927.36	1573			
Joint model	1486.98	1758			
<b><u>CSA versus FRCR1</u></b>					
CSA Equated score	503.22	185	4	<i>F(4,889)=2.19</i>	2.38
FRCR1	92.59	159			
Joint model	601.68	897			

The results suggest that the CSA Equated model was statistically significantly different from the AKT model and the SCEs model (which also included Cardiology), but was not statistically significantly different from the FRCR1 model.

### **6.3 THE ANNUAL REVIEW OF COMPETENCE PROGRESSION ('ARCP')**

The Annual Review of Competence Progression is an assessment applicable to doctors in Core Medical Training and Higher Specialty Training. Its purpose is to verify if a trainee's progress in expected areas is appropriate. ARCP is meant to be a formative assessment supporting a trainee's development, rather than serving summative purposes. It is made on a standard scale of outcomes 1 to 9, as described in more detail in section 2.4.6. For the purposes of this research the outcomes were classified into satisfactory progress (outcomes 1 and 6) and unsatisfactory progress (outcomes 2, 3, 4 and 5). Outcomes 7, 8 and 9 were omitted entirely as they indicate situations beyond the standard training; for example, being out of programme for research, or undertaking additional training. Based on the division into two classes of outcomes doctors were assigned to two groups: those who had only satisfactory outcomes throughout training, and those who had at least one unsatisfactory. This resulted in creating a final binary Overall Progress Assessment variable,

which was assumed to be a measure of general performance. The analyses were performed only on those trainees who were in Specialty Training, so that their ARCP performance was not MRCP(UK) related. As previously mentioned, during CMT the ARCP outcome is dependent on whether a trainee passes consecutive parts of MRCP(UK).

### **6.3.1 Descriptive statistics**

The analysed final file contained records of 5,644 trainee doctors. Among them 2,928 were female doctors (51.9%) and 2,713 were male (48.1%); information on sex was missing for three trainees. The majority of the trainees were of other ethnicity than white (52.3%) and were UK graduates (71.3%).

The file contained information on the latest stage of training at which ARCP assessment was made. Nearly 51% (2,884 doctors) had their last assessment during Specialty Training ('ST') in years 1 to 6; 45% (2,514 doctors) had their last assessment made during Core Medical Training ('CMT') in years 1 to 3. The last 4% (228 doctors) were in other approved forms of training such as Out of Programme ('OOP') or Locum Appointment Training ('LAT'). Only 18 doctors were in Fixed Term Specialty Training Appointments ('FTSTA').

Due to the reasons explained in the Methodology section, the records of doctors who were still in the CMT or those whose penultimate assessment was obtained during CMT had to be excluded from the analysed sample. Elimination of the first group of the CMT trainees resulted in 3,130 valid cases; the elimination of the second group of the ST trainees resulted in the final sample of 2,979 cases. In the final sample the majority of the trainees had a record of only one assessment (70.5%), or just two ARCP assessments (17.9%). Larger numbers of assessments in the programme were less prevalent as presented in Figure 26. The analyses on groups of doctors who had more than two assessments could have been contaminated by their CMT training assessments.

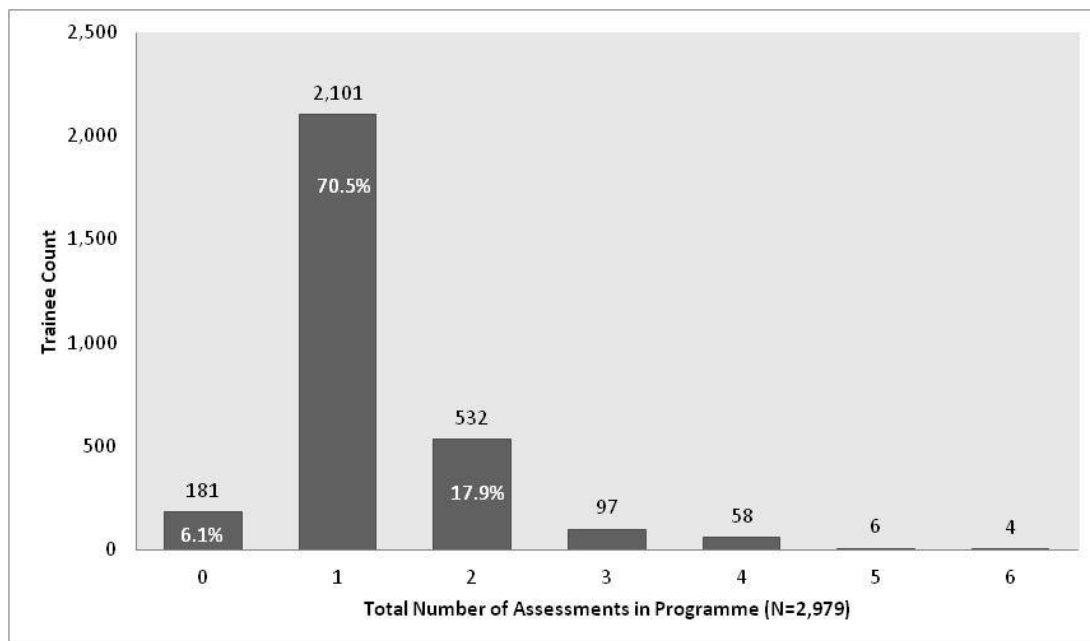


Figure 26. Frequency of the number of assessments in the analysed ARCP sample ( $n=2,979$ ).

The distribution of the overall number of assessments was skewed for two major reasons. First, the dataset contained assessments from years 2010 and 2011 only, which meant that there was a limited number of assessments that the trainees were obliged to participate in (and the required standard is once a year). Secondly, this dataset was affected by censoring as ARCP is a continuous process, meaning that those who joined the training programme later, for example in 2011, had fewer assessments or may have had none.

If a candidate had more than one record in the dataset only the final record was taken into account, as it presented the most complete picture of the trainee's performance during training. Based on the Overall Progress Assessment, there were 2,142 candidates with all satisfactory outcomes (76.6%) and 656 candidates with at least one unsatisfactory outcome (23.4%). One hundred and eighty one candidates remained unassessed by the closing date of the file. The distributions between those who had all satisfactory and at least one unsatisfactory outcome varied depending on the total number of assessments in the programme. This is presented in Figure 27 (% of an overall number of candidates within each group).

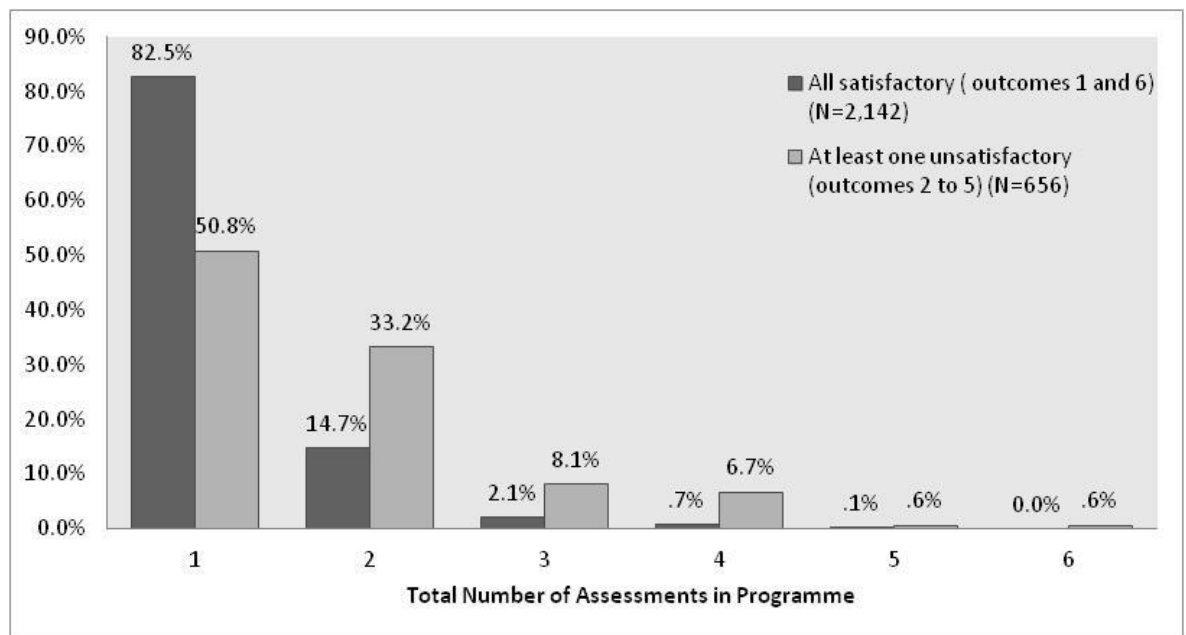


Figure 27. Frequencies of the number of assessments in groups based on the Overall Progress Assessment (All satisfactory versus At least one unsatisfactory).

There were more trainees (% value of the total in the group) with only one assessment who were in the Satisfactory Progress group, while in the Unsatisfactory Progress group there were significantly more trainees with two assessments and more. The explanation for these differences may lie in administrative procedures, as those trainees who get an unsatisfactory outcome in an assessment are often required to submit themselves to another review within 3 to 6 months, which effectively means they have more records of assessments in general.

### 6.3.1.2 MRCP(UK) distributions

The effective sample sizes on which the analyses were performed were even smaller than previously mentioned ( $n=2,979$ ), due to censoring of the MRCP(UK) dataset. There were 2,539 valid cases for Part I results, with 2,836 cases for Part II results, and 2,916 records of PACES results. The distributions of MRCP(UK) parts results were tested for normality using a K-S test, and they indicated that only Part II results were normally distributed ( $Z_{KS} = 0.93$ ,  $p=0.356$ ) in that sample. Part I and PACES distributions were significantly different from normal ( $Z_{KS} = 1.64$ ,  $p=0.009$ ;  $Z_{KS} = 2.66$ ,  $p<0.001$ , respectively). Despite non-normality of the distributions the analyses were performed using parametric methods, as previously argued for in Chapter 4.

### 6.3.2 Inferential statistics

#### 6.3.2.1 Contrasting Groups

In order to test if a group of MRCP(UK) Highfliers was performing better than the rest of the trainees, as it was in the case of all previous criteria, a comparison of means with the independent samples t-test was performed on the first attempt scores in Part I, Part II, and PACES with Overall Progress Assessment being the factor. Two groups were compared: of those with only satisfactory outcomes *versus* those who had at least one unsatisfactory.

Table 63 (below) summarizes the number of cases used in the analysis, the mean scores, and the values of the t-tests with associated *p*-values.

**Table 63. Comparisons of mean scores in MRCP(UK) parts (independent samples t-tests) with Overall Progress Assessments as the factor.**

<i>Group</i>	<i>Valid cases</i>	<i>Mean and SD</i>	<i>Independent samples t-test</i>	<i>Effect size (r)</i>
<b><u>Part I</u></b>				
Satisfactory Progress	1,839	1.45 (SD=9.35)	$t(2,379)=8.05, p<0.001$	0.19
Unsatisfactory Progress	542	-2.26 (SD=9.63)		
<b><u>Part II</u></b>				
Satisfactory Progress	2,058	4.66 (SD=6.85)	$t(2,659)=9.04, p<0.001$	0.21
Unsatisfactory Progress	603	1.79 (SD=6.84)		
<b><u>PACES</u></b>				
Satisfactory Progress	2,115	0.15 (SD=6.25)	$t(2,736)=6.04, p<0.001$	0.14
Unsatisfactory Progress	623	-1.58 (SD=6.37)		

There was a significant difference between the mean scores in all three parts of MRCP(UK) between the two groups of the Overall Progress Assessment. Doctors with only satisfactory outcomes throughout training scored significantly higher at their first attempt in all three parts of MRCP(UK), in comparison to those who had at least one unsatisfactory outcome. The associated effect sizes were moderately small.

The consistency of these effects was tested within groups with different numbers of recorded assessments also using independent samples t-tests. The results have shown that among trainees with only one assessment record, those who had all satisfactory outcomes scored significantly higher in all three parts of MRCP(UK) than those with at least one unsatisfactory (Part I:  $t(1,819)=7.44$ ,  $p<0.001$ ; Part II:  $t(2,002)=7.69$ ,  $p<0.001$ ; PACES:  $t(2,050)=5.53$ ,  $p<0.001$ ). Among those who had two recorded assessments, the significant differences were observed in Part II first attempt scores only ( $t(505)=2.50$ ,  $p=0.013$ ), with those who had only satisfactory outcomes scoring higher. The groups of trainees with more than two assessments were significantly smaller than the first two, from eighty for the group with three assessments to just three with six assessments. Among trainees with three recorded assessments those with all satisfactory outcomes scored higher in Part I than the other group ( $t(83)=2.17$ ,  $p=0.033$ ), and the analyses on even smaller groups were omitted. Due to the fact that the group sizes declined, analyses based on candidates with three assessments were bootstrapped. Bootstrapping showed that there were no significant differences for candidates with three assessments. Table 64 (next page) summarizes the numbers of valid cases for each group and mean scores for the above-described differences.

The means indicate that trainees with all satisfactory ARCP outcomes in general scored higher than their colleagues who had at least one unsatisfactory outcome during their training.

#### **6.3.2.2 Regression models**

A logistic multiple regression model was fitted (entry method) to the data in order to indicate the relative importance of the MRCP(UK) parts as predictors on worse performance in training, as indicated by the Overall Progress Assessment variable. The model included MRPC(UK) parts as predictors, together with demographic characteristics, namely: sex, PMQ, ethnicity, and age.

The model converged at fourth iteration and was considered well-fitted based on the Hosmer-Lemeshow statistic, which was non-significant ( $\chi^2(8, n=2,326) = 5.77$ ,  $p=0.673$ ). The summary of the model is presented in Table 65.

**Table 64. Comparisons of mean score in MRCP(UK) with Overall Progress Assessments as the factor for groups with different number of assessments**

	<i>Part I</i>		<i>Part II</i>		<i>PACES</i>	
	<i>Satisfactory Progress</i>	<i>Unsatisfactory Progress</i>	<i>Satisfactory Progress</i>	<i>Unsatisfactory Progress</i>	<i>Satisfactory Progress</i>	<i>Unsatisfactory Progress</i>
<b><u>One assessment only</u></b>						
N	1,533	288	1,697	307	1,743	309
Mean	1.55 (SD=9.38)	-2.96 (SD=9.75)	4.72 (SD=6.80)	1.48 (SD=6.78)	0.27 (SD=6.27)	-1.88 (SD=6.49)
<b><u>Two assessments</u></b>						
N	254	173	304	203	312	211
Mean	0.83 (SD=9.28)	-0.67 (SD=9.48)	4.27 (SD=7.02)	2.67 (SD=7.16)	-0.48 (SD=6.08)	-1.23 (SD=6.13)
<b><u>Three assessments</u></b>						
N	41	44	43	47	44	51
Mean	0.07 (SD=9.03)	-4.31 (SD=9.52)	4.01 (SD=7.58)	1.92 (SD=6.49)	-0.59 (SD=6.47)	-1.29 (SD=6.38)
	(bias=-0.07, SE=1.44)	(bias=-0.05, SE=1.37)	(bias=-0.01, SE=1.09)	(bias=-0.03, SE=0.98)	(bias=-0.04, SE=0.94)	(bias=-0.01, SE=0.83)
	95%CI [-2.75, 2.94]	95%CI [-7.15, -1.65]	95%CI [1.94, 6.07]	95%CI [0.04, 3.80]	95%CI [-2.57, 1.27]	95%CI [-2.90, 0.32]



**Table 65. Summary of the logistic regression model for ARCP Unsatisfactory Progress as dependent variable with MRCP(UK) parts and demographic characteristics as predictors.**

<i>Predictor</i>	<i>Odds ratio</i>	<i>Significance</i>	<i>95% CI</i>		<i>Wald statistic</i>
			<i>Lower</i>	<i>Upper</i>	
Part I 1 <sup>st</sup> Attempt result (Z score)	0.85*	0.013	0.75	0.97	6.22
Part II 1 <sup>st</sup> Attempt result (Z score)	0.74**	0.000	0.65	0.85	18.72
PACES 1 <sup>st</sup> Attempt result (Z score)	0.94	0.322	0.83	1.06	0.98
Sex (female)	0.65**	0.000	0.53	0.81	15.20
Ethnicity (white)	1.03	0.805	0.81	1.31	0.06
PMQ (UK grads)	1.02	0.884	0.79	1.32	0.02
Current Age	0.33	0.128	0.08	1.37	2.31

Cox and Snell pseudo-R<sup>2</sup> = 4%; \* significant with  $p < 0.05$ , \*\* significant with  $p < 0.001$

The odds ratios for the model indicate that the lower Part I and Part II scores and being male were all independently predictive of obtaining an unsatisfactory outcome in ARCP. Otherwise the results may be interpreted that higher Part I scores (per one standard deviation), higher Part II scores (per one standard deviation), and being female decreased the likelihood of getting an unsatisfactory result during training. A one standard deviation increase in Part II first attempt scores decreased the likelihood of having an unsatisfactory result in training by 26%, while an increase in Part I scores by each standard deviation decreased the likelihood of having an unsatisfactory result by 15%. Being female decreased the likelihood of having an unsatisfactory result by 35%. Based on the Wald statistic the most important predictor was the first attempt Part II score, followed by sex, and finally followed by Part I scores. The Wald statistic is significantly higher for Part II than for Part I, and considering that standard errors of the odds ratios for these two predictors are of similar size<sup>10</sup>, the effect is driven purely by the strength of the coefficients (odds ratios) associated with each part.

## 6.4 REGISTRATION STATUS BASED ON THE LIST OF REGISTERED MEDICAL PRACTITIONERS

The List of Registered Medical Practitioners is an official register administered by the GMC. Registration with the GMC is mandatory to all doctors who wish to practise medicine in the

<sup>10</sup> The Wald statistic is calculated as squared odds ratio divided by associated standard error. If any of the estimated effects is associated with a particularly large SE, then comparison of two values of Wald statistics may not be effective.

UK. Registration is not equivalent to holding a licence to practise. Detailed information on the differences in registration types is provided in section 2.4.7. Based on the registration status of doctors, three key variables were extracted: Licence Issues, Voluntary Erasure and Administrative Erasure. The Licence Issues collectively refer to any situation where during the course of an investigation by the GMC either any limitation on a licence of a doctor was imposed or when such licence was revoked. 'Voluntary Erasure' refers to the situation where a doctor has decided to relinquish their licence. This status was considered as a measure of underperformance as sometimes doctors do voluntarily decide to erase themselves from the register in the course of an investigation by the GMC. Administrative Erasures are usually a reflection of a doctor failing to meet certain administrative requirements. The first two measures were assumed to be representations of clinical underperformance, while the last measure was considered a distant proxy of conscientiousness (McLachlan *et al.*, 2009; McLachlan, 2010), and hence, of professional attitudes (Finn *et al.*, 2009).

#### **6.4.1 Descriptive statistics**

In the analysed merged dataset (for description see Chapter 3, section 3.2.7) there were 33,359 doctors, of whom 51.9% were male (17,836 cases) and 48.1% were female (16,528 cases); information on sex was missing for five cases. Of the total sample, 64.3% doctors (22,085 cases) were UK graduates. Non-white ethnicity was declared by 57.8% (19,876 cases) of doctors.

Within the sample of 33,359 doctors there were only 330 with Licence Issues (0.99%). Due to censoring of the MRCP(UK) file the analyses were performed on even smaller samples. In the analysis on Licence Issues, among all doctors who had a record of their first attempt in Part I ( $n=25,447$ ) only 220 cases of Licence Issues occurred, 926 cases of Administrative Erasures, and 755 Voluntary Relinquishments. In the case of Part II these numbers were smaller; there were 135 cases of Licence Issues, 654 cases of Administrative Erasures, and 538 cases of Relinquished Licences. In the case of PACES there were 143 records of Licence Issues, 624 Administrative Erasures were recorded, and 521 Relinquished Licences occurred.

As described previously the process of flagging statuses allowed doctors to have more than one flag. For example, if a doctor was first provisionally registered, then fully registered, and then erased from the LRMP for administrative reasons within the fifty-month period the data were collected for, then they would have been associated with three flags in total.

In the analysed file there were 28,962 (86.8%) doctors who had only one flag within that period, 4,134 records (12.4%) of doctors who had two flags. More than 3 flags (and the maximum was five) were recorded for only 259 doctors (0.8%).

In the last group of doctors who had the most flags, 118 were erased for administrative reasons, 17 doctors had their licence revoked after the fitness to practice panel, 55 relinquished their licence, 143 decided to be registered without licence, 74 were suspended, 77 had conditions imposed, 32 had undertakings, and 21 had warnings. A total of 111 (42.9%) experienced at least one of the licence limitations, which is consistent with the Fitness to Practice review procedures.

The distributions of Part I, Part II, and PACES first attempt scores were tested for normality and it was established they were significantly different from normal ( $Z_{K-S}=3.42$ ,  $p<0.001$ ,  $Z_{K-S}=2.69$ ,  $p<0.001$ ,  $Z_{K-S}=9.24$ ,  $p<0.001$ , respectively). However, analogously to previous analyses parametric tests were used to analyse the data.

#### **6.4.2 Inferential statistics**

MRCP(UK) first attempt results and demographics were treated as predictors and the three key variables based on the flag system described in Chapter 3, section 3.2.7, were the dependent variables: the Licence Issues, Voluntary Erasure, and Administrative Erasure. First, the univariate tests were calculated to compare the Part I, Part II and PACES first attempt scores between doctors with and without Licence Issues, between doctors who relinquished their Licence and those who did not, and between doctors who were struck off due to administrative reasons versus those who were not. Second, binary multiple logistic regression was used (block entry method) to model the independent relative predictive effects of the three parts of MRCP(UK) and other demographic factors on the three key outcome measures.

##### **6.4.2.1 Contrasting Groups**

###### ***The Licence Issues***

Licence Issues were found to be significantly affected by the first attempt results in Part I, Part II, and PACES ( $p<0.001$ ). Doctors experiencing Licence Issues scored significantly lower in all three parts of the MRCP(UK) in comparison to their colleagues in good standing. The mean scores (bootstrapped), the values of the t-tests, and the effect sizes are presented in Table 66.

**Table 66. Comparison of mean scores in MRCP(UK) parts (independent samples t-tests) for Licence Issues as the factor.**

<i>MRCP(UK) Part</i>	<i>Mean score with SDs</i>		<i>Independent samples t-tests</i>	<i>Effect size (r)</i>
	<i>No Licence Issues</i>	<i>With Licence Issues</i>		
Part I	-1.98 (SD=11.18) bias=-0.003 SE=0.070 95%CI [-2.12, -1.83] n=25,227	-7.62 (SD=11.69) bias=-0.042 SE=0.792 95%CI [-9.19, -6.11] n=220	$t(25,445)=7.45$ , $p<0.001$	0.24
Part II	4.37 (SD=7.77) bias = -0.001 SE =0.057 95%CI [4.25, 4.47] n=18,622	0.97 (SD=7.55) bias = -0.003 SE =0.672 95%CI[-0.35, 2.31] n=135	$t(18,755)=5.07$ , $p<0.001$	0.22
PACES	-1.01 (SD=6.54) bias = 0.002 SE =0.047 95%CI [-1.11, -0.92] n=18,617	-5.28 (SD=6.98) bias = 0.010 SE =0.602 95%CI [-6.44, -4.09] n=143	$t(18,758)=7.77$ , $p<0.001$	0.30

The mean score for the group of doctors with Licence Issues showed a higher error and more bias, although neither should significantly affect the results of the statistical tests. The effect sizes ( $r$ ) were consistent and should be considered moderate in magnitude.

### ***Relinquished Licence***

A highly significant effect of the first attempt scores in PACES ( $p<0.001$ ) was found on voluntary erasures. The means scores, independent samples t-tests results, and effect sizes are presented in the Table 67.

**Table 67. Comparison of mean scores in MRCP(UK) (independent samples t-tests) between doctors with relinquished licences and the rest of the sample.**

<i>MRCP(UK) Part</i>	<i>Mean scores with SDs</i>		<i>Independent samples t-tests</i>	<i>Effect size (r)</i>
	<i>Not relinquished</i>	<i>Relinquished</i>		
Part I	-2.02 (SD=11.21) bias = 0.003 SE =0.070 95%CI [-2.14,-1.87] n=24,692	-2.31 (SD=10.93) bias = 0.008 SE =0.404 95%CI [-3.14, -1.53] n=755	$t(25,445)=0.72$ , $p=0.472$	0.01
Part II	4.35 (SD=7.78) bias =0.001 SE =0.055 95%CI [4.24, 4.47] n=18,219	3.91 (SD=7.33) bias = 0.016 SE =0.317 95%CI [3.24, 4.54] n=538	$t(18,755)=1.32$ , $p=0.186$	0.03
PACES	-1.02 (SD=6.55) bias = ~0.00 SE =0.046 95%CI [-1.11, -0.93] n=18,239	-1.97 (SD=6.64) bias = 0.003 SE =0.276 95%CI [-2.49, -1.46] n=521	$t(18,758)=3.27$ , $p=0.001$	0.07

The effect sizes were consistent and should be considered extremely small in terms of magnitude. Doctors who relinquished their Licence scored significantly lower in PACES than other doctors in the sample. A similar effect of significance of PACES ( $p<0.001$ ) was observed after removing from the dataset those doctors who were already flagged as experiencing Licence Issues; the procedure excluded six records in Part I results (leaving 749 cases), three records in Part II results (leaving 535 cases), and three records in case of PACES results (518 cases).

### ***Administrative issues***

It was established that the group of doctors who were removed from the register for administrative reasons scored on average significantly lower in all three parts of MRCP(UK) than their colleagues who did not have their licence erased. Those differences were highly significant ( $p<0.001$ ), and the effect sizes were moderate in magnitude. The mean scores, values of the t-tests, and the effect sizes are presented in Table 68.

**Table 68. Comparison of mean scores in MRCP(UK) parts (independent samples t-tests) results between doctors erased for administrative reasons and not erased.**

<i>MRCP(UK) part</i>	<i>Mean Scores with SDs</i>		<i>Independent samples t-test</i>	<i>Effect size (r)</i>
	<i>On the register</i>	<i>Admin Erasure</i>		
Part I	-1.88 (SD=11.20) bias =~0.00 SE =0.073 95%CI [-2.03, -1.74] n=24,521	-5.79 (SD=10.38) bias = -0.010 SE =0.334 95%CI [-6.46, -5.16] n=926	$t(25,445)=10.43$ , $p<0.001$	0.18
Part II	4.44 (SD=7.78) bias =-0.002 SE =0.056 95%CI [4.32, 4.54] n=18,103	1.70 (SD=6.89) bias =-0.009 SE =0.274 95%CI [1.15, 2.23] n=654	$t(18,755)=8.86$ , $p<0.001$	0.18
PACES	-0.93 (SD=6.52) bias =~0.00 SE =0.046 95%CI [-1.02,-0.84] n=18,136	-4.41 (SD=6.67) bias = -0.011 SE =0.264 95%CI [-4.92,-3.88] n=624	$t(18,758)=13.11$ , $p<0.001$	0.26

In the dataset there was no overlap between the groups of those who experienced administrative problems and the Licence Issues. However, the analyses excluded doctors who relinquished their licence, i.e. twelve cases from the Part I sample (leaving 914 cases), nine cases from Part II sample (leaving 645 cases), and eight cases from the PACES sample (leaving 616 cases). The findings were analogous and the differences continued to be highly significant ( $p<0.001$ ).

#### **6.4.2.2 Regression models**

The multiple logistic regression models were fitted based on the 12,199 candidates who had full records of Parts I, II and PACES scores. First attempt results at parts of MRCP(UK) and demographic characteristics (sex, current age, ethnicity, and PMQ) were entered as predictors into each of the models (block entry method).

#### **Licence Issues**

The logistic regression model for Licence Issues converged at the ninth iteration. It should be regarded as well-fitted based on the Hosmer-Lemeshow statistic ( $\chi^2(8, n=12,199) = 9.19$ ,  $p=0.327$ ). Table 69 summarizes the results of the logistic regression for all predictors.

**Table 69. Summary of the logistic regression model for occurrences of Licence Issues as dependent binary variable with MRCP(UK) parts and demographic characteristics as predictors.**

<i>Predictor</i>	<i>Odds ratio</i>	<i>Significance</i>	<i>95% Confidence Interval</i>		<i>Wald statistic</i>
			<i>Lower</i>	<i>Upper</i>	
Part I 1 <sup>st</sup> Attempt result (Z-score)	1.12	0.513	0.80	1.58	0.43
Part II 1 <sup>st</sup> Attempt result (Z-score)	0.88	0.448	0.64	1.22	0.58
PACES 1 <sup>st</sup> Attempt result (Z-score)	0.73*	0.020	0.56	0.95	5.38
PMQ (UK graduates)	1.41	0.298	0.74	2.70	1.08
Ethnicity (white)	0.78	0.411	0.44	1.40	0.68
Sex (female)	0.28**	0.000	0.15	0.52	16.2
Current Age	1.07*	0.016	1.01	1.14	5.80

Cox and Snell pseudo-R<sup>2</sup> = 0.4%, \* significant with  $p < 0.05$ , \*\* significant with  $p < 0.001$

In the order of relative importance, male sex (odds ratio = 0.28, 95%CI [0.15, 0.52],  $p < 0.001$ ), older age (odds ratio =  $1.07 \pm 0.06$  95%CI,  $p = 0.016$ ) and lower PACES results (odds ratio = 0.73 95%CI [0.56, 0.95],  $p = 0.020$ ) were all significant predictors that increased the likelihood of Licence Issues. Part I and Part II were non-significant when PACES results were taken into account. Similarly, there was no significant effect of UK primary medical qualification or ethnicity.

### ***Voluntary Erasure (relinquished licence)***

The logistic regression model should be considered well-fitted based on the Hosmer-Lemeshow test ( $\chi^2(8, n=12,127)=10.03$ ,  $p=0.263$ ; converged at the seventh iteration). The model is summarised in Table 70 (next page).

A logistic regression model predicting voluntary erasure using the same predictors as for the Licence Issues indicated that higher PACES scores (odds ratio=1.17 95%CI [1.02, 1.34],  $p=0.024$ ) were an independent predictor of taking voluntary erasure, as was having a non-UK primary medical qualification (odds ratio = 0.42 95% CI [0.31, 0.57],  $p < 0.001$ ), and being older in age (odds ratio =  $1.06 \pm 0.03$  95% CI,  $p < 0.001$ ).

**Table 70. Summary of the logistic regression model for Voluntary Erasure as binary dependent variable with MRCP(UK) parts and demographic characteristics as predictors.**

<i>Predictor</i>	<i>Odds ratio</i>	<i>Significance</i>	<i>95% Confidence Interval</i>		<i>Wald statistic</i>
			<i>lower</i>	<i>upper</i>	
Part I 1 <sup>st</sup> Attempt result (Z-score)	1.09	0.309	0.92	1.29	1.04
Part II 1 <sup>st</sup> Attempt result (Z-score)	1.12	0.169	0.95	1.30	1.89
PACES 1 <sup>st</sup> Attempt result (Z-score)	1.17*	0.024	1.02	1.34	5.13
PMQ (UK grads)	0.42**	0.000	0.31	0.56	30.47
Ethnicity (white)	0.82	0.138	0.62	1.07	2.20
Sex (female)	1.02	0.879	0.81	1.29	0.02
Current Age	1.06**	0.000	1.03	1.09	12.54

Cox & Snell pseudo- $R^2$  = 7%; \* significant with  $p < 0.05$ , \*\* significant with  $p < 0.001$

The other predictors were non-significant. Based on the Wald statistic, being a non-UK graduate and being older had more impact on making a decision on voluntary erasure than PACES scores. The model seems to contradict the findings of the univariate statistics, which suggested that voluntary erasure was associated with lower PACES scores. This contradiction might result from introducing demographic factors and all MRCP(UK) parts into one model; however, a separate investigation on the individual reasons for relinquishing licence to practise would be required to make further inferences.

### ***Administrative erasures***

A logistic regression model predicting erasure for administrative reasons was fitted to the data after excluding doctors who experienced Licence Issues and who voluntarily gave up licences ( $n=11,804$ ). The model employed the same predictors as used in previous models. The model should be considered well-fitted based on the Hosmer-Lemeshow test ( $\chi^2(8, n=11,804)=80.01, p=0.433$ ; converged at the seventh iteration). The model is summarised in Table 71.



**Table 71. Summary of the logistic regression model for Administrative Erasures as binary dependent variable with MRCP(UK) parts and demographic characteristics as predictors.**

<i>Predictor</i>	<i>Odds ratio</i>	<i>Significance</i>	<i>95% Confidence Interval</i>		<i>Wald statistic</i>
			<i>Lower</i>	<i>Upper</i>	
Part I 1 <sup>st</sup> Attempt result (Z-score)	1.24*	0.009	1.05	1.45	6.80
Part II 1 <sup>st</sup> Attempt result (Z-score)	1.06	0.489	0.91	1.23	0.48
PACES 1 <sup>st</sup> Attempt result (Z-score)	0.84*	0.008	0.74	0.96	7.04
PMQ (UK grads)	0.35**	0.000	0.26	0.47	46.61
Ethnicity (white)	0.39**	0.000	0.28	0.53	33.82
Sex (female)	0.79	0.062	0.62	1.01	3.48
Current Age	1.03*	0.034	1.00	1.06	4.49

Cox & Snell pseudo-R<sup>2</sup> = 2.4%; \* significant with  $p < 0.05$ , \*\* significant with  $p < 0.001$

The model indicated that apart from Part II scores and sex all predictors were significant. The results show that *higher* Part I scores (odds ratio = 1.23 95%CI [1.05, 1.45],  $p=0.009$ ) and *lower* PACES scores (odds ratio = 0.84±0.12 95%CI,  $p=0.008$ ) were independent predictors of being removed from the register for administrative reasons. In addition, being a UK graduate (odds ratio=0.35 95%CI [0.26,0.47],  $p<0.001$ ), being white (odd ratio=0.39 95%CI [0.28, 0.53],  $p<0.001$ ), and being younger (odds ratio = 1.03±0.03 95%CI,  $p=0.034$ ) decreased the likelihood of being struck off. Based on the Wald statistic, the most important predictors were PMQ, followed by ethnicity, PACES scores, Part I scores, and age.

## 6.5 BEING SUBJECT TO INVESTIGATION BY THE GMC FITNESS TO PRACTICE

### PROCEDURES

Being subject to an investigation by the GMC Fitness to Practice ('GMC FtP') procedures was considered a measure of clinical underperformance on the same grounds as in the case of the LRMP Licence Issues. The list of doctors investigated for unsatisfactory performance at work was obtained from the GMC, and it contained only those cases that were not dismissed in the earlier stages of the GMC review (for details see section 2.4.8).

#### 6.5.1 Descriptive statistics

A list of 820 doctors who were investigated by the Fitness to Practice procedures between 1998 and 2008 was provided. Considering the overall number of doctors who had an LRMP

record being close to 327,000, the reviews are an extremely rare situation; they affected only 0.3% of doctors. On the other hand, cases of identified Licence Issues discussed in section 6.4 were more than four times more prevalent, reaching up to 1.2% of the LRMP overall number of records. The cases of doctors from the GMC FtP list were also found in the LRMP status groups, which is summarised in Table 72.

**Table 72. Overlap between GMC FtP list of investigated cases and groups of doctors with selected LRMP registration statuses.**

<i>Registration Status</i>	<i>Count (common cases)</i>
Administrative Reasons	56
Erased after FtP procedure	124
Relinquished Licences	171
Suspended	155
Warnings Issued	26
Undertakings	85
Licence Issues (jointly)	349

The only information extracted from the GMC file was a binary variable denoting if a person was or was not reviewed by the panel. This information was merged with the History File dated May 2013, in which there were 47,759 cases with Part I first attempt results, 30,910 cases of Part II first attempt results, and 27,271 cases of first attempt PACES results. The varied number of cases depending on MRCP(UK) part was a result of data censoring.

The merge resulted in matching only five cases of doctors who went through the panel and had a record of first attempt Part I scores, six cases with first attempt Part II scores, and eight cases with first attempt PACES scores. The most likely reason for the low number of cases was that the full Fitness to Practice Panel review is quite uncommon, which might have been emphasized by the short overlap time between the History File and the GMC File. The GMC file contained information on panels taking place from 1998 to 2008, while the History File contained exam results from 2003 onwards, which makes an overlap of six years. Although it is possible for the doctors to be referred as soon as they start the Foundation Programme, the likelihood of being referred to the panel increases with years of practice (Humphrey, Hickman, & Gulliford, 2011; Wakeford, 2011). This makes referrals for candidates in the History File, who are at the early stages of their career, relatively unlikely.

The analyses were carried out, but due to the small sample size their results are presented mainly for qualitative rather than quantitative purposes.

### **6.5.2 Contrasting Groups**

A univariate comparison of means analysis was performed between two groups: those who were on the GMC FtP procedures list and those who were not. Independent samples t-tests were employed, which were subsequently bootstrapped with stratification. The results indicated that doctors who were not reviewed by the GMC FtP procedures scored significantly higher than their GMC-listed colleagues.

In the case of Part I scores the difference reached 14% and was highly significant ( $t(47,757) = 2.67, p=0.008, r=0.60$ ), with those reviewed scoring on average 17.33% below the pass-mark on their first attempt and the rest of the sample scoring on average 3.09% below the pass-mark. Bootstrapping confirmed the significance of the observed difference ( $p=0.001$ ). The difference in Part II scores was smaller, i.e. 6.2%, and almost significant ( $t(30,908) = 1.84, p=0.066, r=0.44$ ); however, when bootstrapped the difference became highly significant ( $p=0.001$ ). Doctors who were reviewed by the GMC procedures scored on average 2.3% below the pass-mark, while the average in the second group was 3.92% (above the pass-mark). The difference was even smaller in the case of PACES results; 4.96 % ( $t(27,269)=2.03, p=0.042, r=0.33$ ). Bootstrapping further confirmed the significance ( $p=0.026$ ) of the obtained results. Doctors who went through the GMC FtP procedures scored on average 6.97% below the pass-mark, while the rest of the doctors in the dataset scored on average 1.99% below the pass-mark.

The results seem to suggest that Part I, Part II, and PACES scores are predictive of being reviewed by the GMC FtP procedures, with a predominant role played by Part I scores.

## **SUMMARY AND DISCUSSION**

The purpose of this chapter was to test the hypothesis that MRCP(UK) predicts clinical skills and attitudes of doctors. This hypothesis was supported by the findings.

Firstly, it was consistently found by means of univariate methods that there were significant differences in all three parts of MRCP(UK) between doctors in good standing and doctors who experienced any issues in their training or clinical practice. The records of the progress assessment during the Specialty Training ('ARCP') showed that doctors with satisfactory progress throughout training scored significantly higher in all three parts of MRCP(UK) than those with at least unsatisfactory assessment outcome. Further, doctors whose licence was

in any way temporarily limited or revoked scored significantly lower in all three parts of MRCP(UK); this was the same for those doctors who had their licences erased for administrative reasons. The results from a small group of doctors who were on the list of cases investigated by the GMC FtP procedures, and had records in the History File, supported the findings obtained from the LRMP dataset. Doctors whose performance was investigated had significantly lower scores than doctors whose performance was not questioned, with a particularly large discrepancy observed for Part I scores. Although these findings were statistically significant, they were obtained based on a very small sample and should therefore be treated as a rather qualitative support to the more robust findings based on the LRMP records. All of the above results generally indicate that doctors who scored higher in MRCP(UK) performed better in clinical situations and were less likely to have issues with registration. Similar results were obtained by Papadakis *et al.* (Papadakis *et al.*, 2005; Papadakis *et al.*, 2008); lower certification scores and lower MCAT scores were associated with higher risk of a disciplinary procedure. A study by Teherani and colleagues (Teherani, Hodgson, Banach, & Papadakis, 2005) has shown that disciplinary actions resulted mainly from poor personal reliability and responsibility, lack of improvement, and lapses in motivation. It would only be logical that these broad personal characteristics affected not only clinical performance, but also particular life events such preparation for or performance at an examination. Therefore, the relationship between exam scores (e.g. knowledge exams) and poor clinical performance leading to disciplinary actions seems not to be straightforward, but rather be mediated by personal or temporal characteristics.

Secondly, there was a statistically significant relationship between MRCP(UK) and clinical examination scores. The MRCP(UK) Highfliers scored significantly better in CSA irrespective of the marking scheme when compared to Typical candidates. Similarly, the FRCF1 Highfliers scored significantly higher in the FRCR2 clinical examination. The uncorrected correlation coefficients between MRCP(UK) parts and CSA scores were of moderate strength ranging from 0.32 to 0.58 (Pearson's *r*). However, the coefficients between PACES and CSA, irrespective of the marking scheme, were always higher than those associated with Part I and Part II scores. These results suggest that although a knowledge component is present in the CSA clinical skills examination, it mainly measures skills similar to those assessed by PACES. This is in concordance with what is known about the two exams, which are similar both in contents and form. The multivariate analysis of MRCP(UK) parts onto CSA scores found that PACES was the most important predictor when taking into account the effect of Part I and Part II scores. Irrespective of the CSA marking scheme the

standardised beta coefficient for PACES was always at least twice as high as for the other two parts. This pattern supported conclusions from Chapter 5 on the similarity of the constructs and forms of the exams. The results of the analyses on FRCR2 clinical examination were less conclusive. A significant albeit small correlation was found between FRCR2 clinical scores and Part II with the coefficient equal to  $r=0.21$ . The correlation between PACES and FRCR2 was very small ( $r=0.08$ ) and non-significant. Both these results suggest that the FRCR2 clinical examination may be more focused on knowledge than on clinical skills, which is supported by the size of the non-significant coefficient with Part I ( $r=0.20$ ). Both PACES and FRCR2 clinical are clinical examinations that include stations with cases of clinical scenarios; therefore, the probable cause for low correlation between them was likely not to result from the incongruence of the forms. At the same time it is probable that skills required of a clinical oncologist are extremely field-specific and the FRCR2 clinical examination aims to test those skills in particular. This means that only generic clinical skills would constitute the common pool of tested abilities for both PACES and FRCR2. Such a pattern would be consistent with the difference in constructs and would be reflected in a weak correlation between the scores of both exams, as observed.

The multivariate analyses on the measures of clinical performance, such as the satisfactory progress throughout Specialty Training, experiencing Licence Issues, and being investigated by the GMC shed light on the relative influence of each of the parts on measures of on-the-job performance. These binary logistic models were, however, not conclusive in terms of the predominant role of PACES results over Part I and Part II for these outcome variables, as was observed in the case of the linear regression model fitted for the CSA examination. The model describing the effect of MRCP(UK) parts on assessment of progression through training indicated that Part I scores and Part II scores were better predictors than PACES. It may be argued that the training stage assessments still focus on doctors acquiring mainly knowledge rather than clinical skills, or that the evidence of acquired knowledge is more prevalent or more feasible to provide during the time when assessments are made. On the other hand, the relative importance of PACES over the written Parts I and II was confirmed in the Licence Issues logistic regression model, which indicated that lower PACES scores were predictive of experiencing trouble with licencing. These findings seem to suggest that clinical skills are crucial during an on-the-job performance assessment process, which validates PACES as a measure of clinical skills.

PACES scores were also a predictor for Voluntary Erasures from the LRMP, yet the results obtained with univariate and multivariate analyses give contradictory effects. The

univariate analyses showed that doctors who decided to relinquish their licence obtained lower scores in PACES, while a multivariate model suggests that Voluntary Erasure was predicted by higher PACES scores. Voluntary Erasure is sometimes used by doctors investigated by the GMC as a way of avoiding forced erasure; therefore, voluntary erasures might suggest lower performance. However, the number of voluntary erasures in the LRMP is much higher than the number of doctors who relinquished their licence during the FtP procedures, which may mean there are other reasons for relinquishing a licence; reasons that are not performance related. Whilst speculative, it is probable that, for example, a non-UK trained doctor may decide to leave the UK and return to practise in their home country, or that an excellent practitioner decides to work in research or become an academic teacher. Therefore, the subsequent multivariate analyses also included demographic factors. The results indeed showed that apart from PACES scores, non-UK PMQ, and older age were also significant predictors of Voluntary Erasures. Therefore, the relationship between Voluntary Erasures, poor performance, and MRCP(UK) parts was found to be more complex than assumed. Without a further study inferences relating relinquishment to poor performance are not sufficiently substantiated. In fact, personal reasons behind relinquishment would constitute an interesting area to investigate within a qualitative study, which could be a subject for future research.

The multivariate analyses of the relationship between MRCP(UK) parts and being erased from the LRMP for administrative reasons showed that Part I and PACES were both significant predictors, together with several demographic factors included in the model. Being erased for administrative reasons was assumed to be a proxy measure of conscientiousness (Finn *et al.*, 2009; McLachlan *et al.*, 2009), as it happens mainly when the annual membership fees are not paid on-time, or when a doctor fails to provide documentation required by the GMC. Since MRCP(UK) is meant to measure professional attitudes, Administrative Erasures should consistently correlate with all MRCP(UK) parts. The fact that administrative erasures were not predicted by Part II scores in presence of Part I and PACES requires further investigation. Anecdotally, it could be argued that Part II is in some way less susceptible to variability in conscientiousness than the other MRCP(UK) parts.

In general, the results presented in Chapter 6 support the hypothesis and the predictive validity of MRCP(UK), despite the fact that the pattern of relationship in the case of clinical performance measures is not as straightforward as in the case of knowledge exams. Clinical and on-the-job performance are complex constructs dependent on many factors, many of

which were not included in this research. A separate study could potentially focus solely on identifying such factors and investigating their influence on clinical performance, in a systematic manner as done by Ferguson *et al.* (2002) for success in medical school. However, the link between MRCP(UK) scores and clinical performance measures should not be underestimated. The first attempt scores consistently differentiated doctors across measures, and did predict subsequent performance to a certain extent. They also showed that poorly performing doctors were likely to struggle at earlier stages of training, which raises a question of whether their issues could have been resolved sooner, or if such doctors should have passed the exams at all and be licenced. The results, albeit inconsistent, confirm the importance of PACES and potentially other structured clinical assessments as parts of selection procedures and evaluation processes, as they test for skills indispensable in good clinical performance.

It may be argued that the results of this chapter indicate the need to raise the pass-mark in high-stakes examinations to decrease the statistical risk of medical errors and disciplinary hearings. However, there are several reasons why this would probably not be a justified course of action. Firstly, the number of cases of disciplinary investigations and resulting limitations is extremely low – it oscillates around 0.1% to 0.4% of licences issued, depending on the specialty and a limitation. Models presented in this chapter indicate that an increase in a cut-score or a severe limitation to the number of attempts available to candidates could potentially decrease the chance of passing MRCP(UK) by a later-underperforming doctor. However, all models are probabilistic, meaning that it is likely that a higher pass-mark would not eliminate all potentially underperforming doctors. At the same time, it would eliminate from the college those doctors who despite lower MRCP(UK) scores at their first attempt are later performing at an at-least-satisfactory level. With the current demand for medical services, and remembering that the majority of candidates do perform well despite obtaining lower scores in an examination, such a decision would neither be practical nor justifiable.

The findings of this chapter also complemented the results presented in Chapter 5; they confirmed the previously observed pattern of associations between measures representing similar constructs and forms of examination.

In order to estimate the effect of MRCP(UK) parts on the criterion measures jointly, Chapter 7 extends the analyses and presents several meta-analytical models, which include key correlation coefficients and differences of means presented in Chapters 5 and 6.

## Chapter 7

### Meta-analyses

#### ABSTRACT

*Analyses presented in Chapters 5 and 6 provided a variety of results which all suggested that performance in MRCP(UK) predicts subsequent performance in medical exams and in clinical practice. This inference was based on the results of both univariate and multivariate statistics. Among univariate statistics, the correlations and mean differences reflected the relationship between performance in each of the MRCP(UK) parts and the performance in exams and other criterion measures. In order to estimate the average effect of each of the MRCP(UK) parts on the outcome criteria, six meta-analytical models were created. The first three models investigated the effect of MRCP(UK) parts on subsequent examinations; the second group of models investigated the effect of MRCP(UK) parts on measures of underperformance. For the purposes of these analyses, it was assumed that the coefficients associated with a particular MRCP(UK) part were calculated on independent samples. The effect of Part I on subsequent examinations was estimated at 0.69, the effect of Part II at 0.70, and the effect of PACES at 0.48. The effect of MRCP(UK) parts on underperformance criteria was estimated at 0.24 for Part I and at 0.22 for Part II and PACES. The models were highly heterogeneous; however, it was shown that this effect did not result from the sampling error.*

The meta-analyses were used as a means of averaging the estimate of the true effect of the MRCP(UK) parts on the criterion measures. The models were fitted for two groups of effect sizes: correlation coefficients between MRCP(UK) parts and subsequent exam scores, and effect sizes of mean differences between underperforming doctors and those in good standing (converted into point-biserial correlation coefficients). Underperformance was understood as being investigated by the GMC Fitness to Practice Panel, experiencing Licence Issues, Administrative Erasures, Relinquishing Licence, or as an Unsatisfactory Progress throughout training.

#### 7.1 META-ANALYSES OF COEFFICIENTS ASSOCIATED WITH EXAMINATIONS

Correlation coefficients between MRCP(UK) parts and seventeen criteria, which were the exam scores that were presented previously in Chapters 5 and 6, were incorporated into



three separate meta-analytical models, each for a different MRCP(UK) part. The purpose of these models was to estimate the true sizes of the relationships between Part I scores, Part II scores, and PACES scores and the criteria. It is the assumption for the meta-analyses that the effects entered into the models should be obtained from independent samples. However, an issue with the results provided in Chapters 5 and 6 was that there was a certain level of interdependence between the reported results, which was due to an overlap between the samples based on which these correlation coefficients were calculated. Two major problems were identified: one related to the overlap of criterial samples, and the second related to the overlap between MRCP(UK) parts samples.

With respect to the first problem, it was observed that some criterial exams included partial scores, for example, AKT comprised three sub-scores, i.e. AKT Clinical Medicine, Evidence Interpretation, and Organisational Questions, and FRCR2 comprised a written, oral, and clinical exam. The sub-scores and associated correlation coefficients were omitted entirely in the meta-analyses, as an overall score was available for these exams. Secondly, for example, MRCGP CSA and AKT scores were also calculated based on the same sample of candidates, but the outcomes of these two exams were not aggregated; each was a separate criterion. The same was true for the overall FRCR1 score and FRCR2 outcome. In order to be able to enter these exams in the models, the coefficients for FRCR1 and FRCR2 and for CSA and AKT were averaged as weighted means (under FRCR and MRCGP, respectively).

The second issue related to the overlap of samples for the MRCP(UK) parts. For example, the coefficients between MRCP(UK) Part I and CSA, and Part II and CSA, were estimated based on largely the same group of MRCP(UK) candidates. This was true for all MRCP(UK) candidates who had at least two or even all three first attempt scores recorded in the dataset. This would call for averaging the coefficients across MRCP(UK) parts; however, the main interest of this chapter was to estimate separate true effects of the relationships between each part of MRCP(UK) and the criterion measures. For this particular reason, the three parts were considered independent for the purposes of the meta-analyses only. This approach, however, limits the scope of interpretation of the results obtained from the models. As much as the magnitude of the estimates can be compared, in the case of dependent effect sizes the confidence intervals tend to be underestimated (Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013) and should not be compared directly.

The analytical models were fitted using the ‘metafor’ package for *R*. In each case, the models were random-effects models. The analyses used non-standardised raw coefficients corrected for artefacts. Justification for using raw coefficients was already provided in section 3.7.9. The raw coefficients, coefficients corrected for artefacts, standard errors and associated sample sizes are provided in Table 73 on the next page.

The results of the three meta-analyses are summarised in Table 74, where the estimates of the true effects with 95% confidence intervals are provided, together with the variance of the coefficients, and heterogeneity measures.

**Table 74. Summary of three meta-analytical models for coefficients between Part I, Part II and PACES and the criteria associated with examinations.**

<i>Measure</i>	<i>Part I</i>	<i>Part II</i>	<i>PACES</i>
Coefficient Estimate	0.69	0.70	0.48
with 95%CI	95%CI [0.64, 0.75]	95%CI [0.66, 0.74]	95%CI [0.42, 0.53]
Tau <sup>2</sup>	0.008 ( <i>SE</i> =0.004)	0.005 ( <i>SE</i> =0.002)	0.008 ( <i>SE</i> =0.004)
I <sup>2</sup>	85.50%	77.85%	73.15%
H <sup>2</sup>	6.90 (H=2.63)	4.51 (H=2.12)	3.72 (H=1.92)
Cochran’s Q	Q(14)=80.03, <i>p</i> <0.001	Q(14)=61.32, <i>p</i> <0.001	Q(14)=48.75, <i>p</i> <0.001

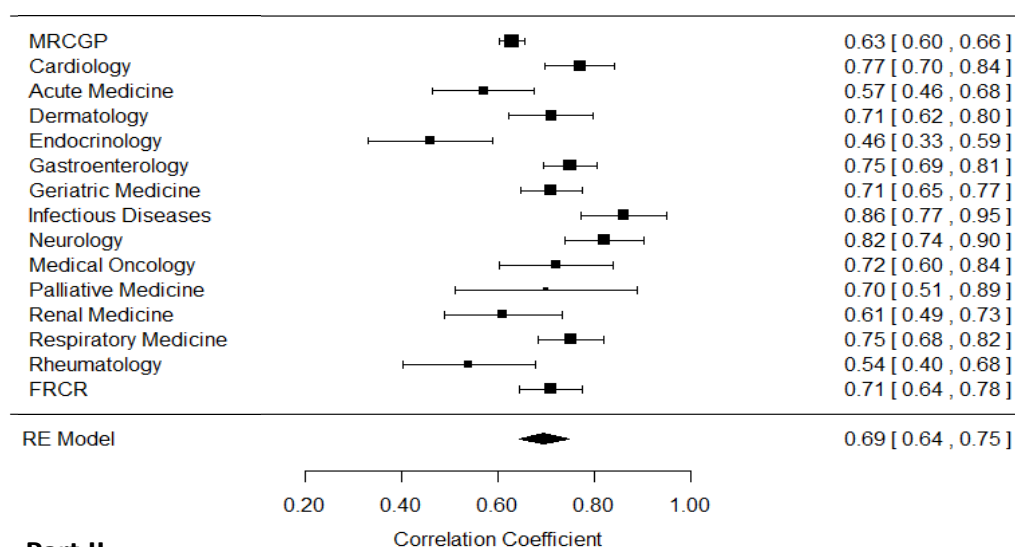
The estimate of the average magnitude of the relationship between Part I scores and criterion measures reached 0.69 with 95%CI [0.64, 0.74]. In the case of Part II, the average effect was estimated at 0.70 with 95%CI [0.66, 0.74], and in the case of PACES, the effect was estimated at 0.48 with 95%CI [0.42, 0.53]. The variances of the estimated effects were close to zero, and based on the values of associated standard errors, they were non-significant. All the models presented in Table 74 should be considered heterogeneous based on the value of Cochran’s Q tests. The heterogeneity seemed to originate from the variation of the true effects (heterogeneity between the studies) rather than sampling errors, as the values of the I<sup>2</sup> statistic were close to or higher than 75%, which is considered large. Based on the value of the H statistic, the total amount of variability in these models was close to two times the amount of sampling variability. This suggests that the relationship between MRCP(UK) parts scores and criterion measures varied across ‘studies’. These effects are visible in Figure 28, where the standard forest plots for Part I, Part II and PACES and the criterion measures are presented.

**Table 73. Raw and corrected Pearson's r coefficients (with SE) between MRCP(UK) parts and criterion measures used in meta-analyses (with sample sizes).**

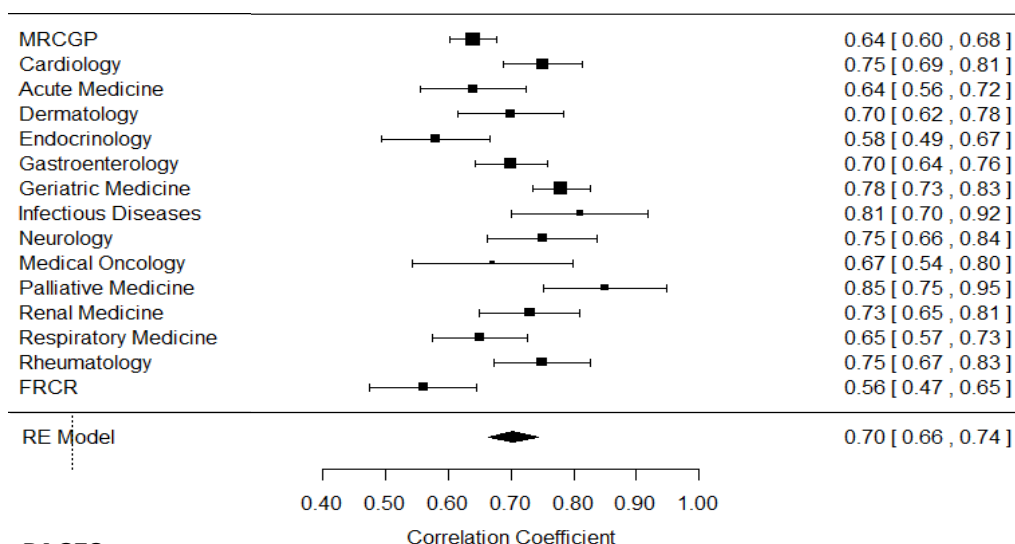
<i>Criterion Measure</i>	<i>Raw coefficients</i>			<i>Corrected coefficients</i>			<i>Sample sizes</i>			<i>Standard Errors (for corrected coefficients)</i>		
	<i>Part I</i>	<i>Part II</i>	<i>PACES</i>	<i>Part I</i>	<i>Part II</i>	<i>PACES</i>	<i>Part I</i>	<i>Part II</i>	<i>PACES</i>	<i>Part I</i>	<i>Part II</i>	<i>PACES</i>
CSA Overall Score Equated*	0.35	0.38	0.46	0.46	0.53	0.59	1,976	938	739	0.020	0.028	0.030
AKT Overall Score*	0.66	0.59	0.43	0.79	0.76	0.53	1,976	938	739	0.014	0.021	0.031
<b>Averaged MRCGP</b>	0.50	0.49	0.45	0.63	0.64	0.56	1,976	938	739	0.017	0.025	0.031
<b>Cardiology</b>	0.53	0.51	0.34	0.77	0.75	0.49	123	184	209	0.058	0.049	0.061
<b>Acute Medicine</b>	0.37	0.48	0.39	0.57	0.64	0.48	159	191	199	0.066	0.056	0.063
<b>Dermatology</b>	0.48	0.52	0.36	0.71	0.70	0.47	125	140	150	0.063	0.061	0.073
<b>Endocrinology</b>	0.29	0.42	0.16	0.46	0.58	0.21	147	225	270	0.074	0.055	0.060
<b>Gastroenterology</b>	0.56	0.53	0.31	0.75	0.70	0.41	235	298	331	0.043	0.042	0.050
<b>Geriatric Medicine</b>	0.49	0.54	0.35	0.71	0.78	0.46	233	282	296	0.046	0.037	0.052
<b>Infectious Diseases</b>	0.66	0.68	0.43	0.86	0.81	0.51	34	39	43	0.090	0.096	0.134
<b>Neurology</b>	0.62	0.62	0.50	0.82	0.75	0.57	63	97	126	0.073	0.068	0.074
<b>Medical Oncology</b>	0.55	0.48	0.28	0.72	0.67	0.34	65	72	83	0.087	0.089	0.104
<b>Palliative Medicine</b>	0.45	0.66	0.48	0.70	0.85	0.68	29	31	32	0.137	0.098	0.134
<b>Renal Medicine</b>	0.45	0.60	0.28	0.61	0.73	0.35	103	132	153	0.079	0.060	0.076
<b>Respiratory Medicine</b>	0.56	0.49	0.41	0.75	0.65	0.53	163	220	235	0.052	0.051	0.056
<b>Rheumatology</b>	0.42	0.63	0.47	0.54	0.75	0.56	105	125	137	0.083	0.060	0.071
FRCR1 Mean Module Mark*	0.52	0.41	0.36	0.74	0.59	0.52	224	249	253	0.045	0.051	0.054
FRCR2 Pass/Fail Score*	0.36	0.32	n.s.	0.62	0.51	n.s.	58	133	246	0.105	0.075	n/a
<b>Averaged FRCR</b>	0.48	0.38	0.36	0.71	0.56	0.52	224	249	253	0.047	0.053	0.054

\*Coefficients for AKT and CSA were averaged to Averaged MRCGP, and those for FRCR1 and FRCR2 coefficients were averaged to Averaged FRCR (weighted means).

## Part I



## Part II



## PACES

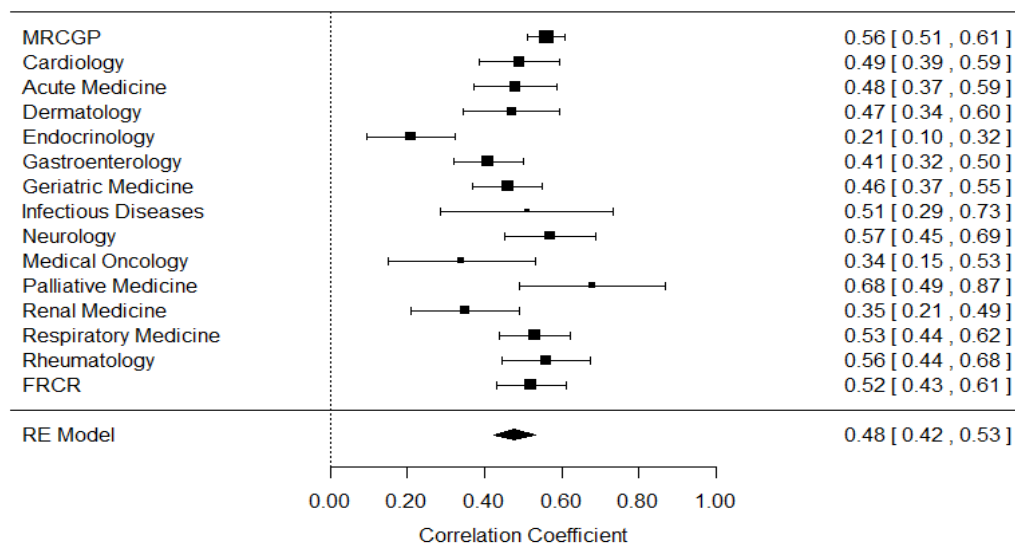


Figure 28. Standard forest plots for effects of MRCP(UK) parts on examinations criteria.

Inspection of the funnel plots for the three meta-analyses showed no significant asymmetry, suggesting there was no bias in the way the coefficients were chosen for the meta-analyses, effectively suggesting no researcher bias. Figure 29 (below) shows only the funnel plot for Part I. Due to repetitiveness of information presented on the funnel plots, the plots for other models are provided in Appendix E.

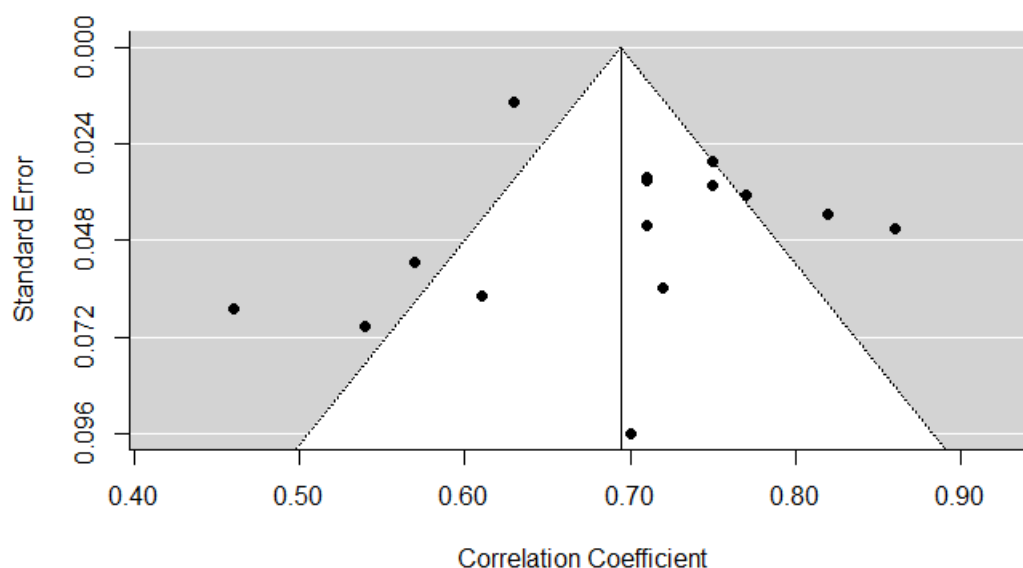


Figure 29. Funnel plot for meta-analytical model on coefficients associated with Part I and examinations.

## 7.2 META-ANALYSES OF GENERAL UNDERPERFORMANCE EFFECT SIZES

Similar to the analyses performed on coefficients associated with examinations, three meta-analytical models were fitted for the effect sizes obtained during analyses on underperformance. There were only five such studies, namely, investigation by the GMC Fitness to Practice Panel, Licence Issues, Administrative Erasures, Relinquished Licences, and Unsatisfactory Progress during training. These effect sizes were not corrected for any artefacts. Table 75 on the next page contains the point-biserial correlation coefficients, together with the associated standard errors and degrees of freedom for the original independent samples t-test statistics. The effect sizes are provided in absolute values; however, MRCP(UK) scores and underperformance showed an inverse relationship – the better the scores in MRCP(UK), the less likely the occurrence of underperformance.

**Table 75. Effect sizes for underperformance related criteria for all MRCP(UK) parts (with SE and degrees of freedom).**

<i>Measure</i>	<i>Effect sizes (r)</i>			<i>Standard Errors for Coefficients</i>			<i>Degrees of Freedom*</i>		
	<i>Part I</i>	<i>Part II</i>	<i>PACES</i>	<i>Part I</i>	<i>Part II</i>	<i>PACES</i>	<i>Part I</i>	<i>Part II</i>	<i>PACES</i>
ARCP Unsatisfactory Progress	0.19	0.21	0.14	0.039	0.037	0.037	2,379	2,659	2,736
Licence Issues	0.24	0.22	0.30	0.012	0.014	0.013	25,445	18,755	18,758
Relinquished Licences	0.01	0.03	0.07	0.012	0.014	0.014	25,445	18,755	18,758
Administrative Erasures	0.18	0.18	0.26	0.012	0.014	0.013	25,445	18,755	18,758
Investigation by the FtP	0.60	0.44	0.33	0.006	0.009	0.011	47,757	30,908	27,269

\*Degrees of freedom correspond to the sample size based on which a Pearson's *r* coefficient would have been calculated

The effect sizes for Relinquished Licences were extremely small, suggesting that the relationship between MRCP(UK) scores and giving up registration was close to zero. As it was previously argued reasons for relinquishing a licence require further investigation; however, since a significant effect was observed for PACES the three meta-analyses incorporated the associated effect sizes. These models are summarized in Table 76 (below).

**Table 76. Summary of the meta-analytical models for measures of underperformance related to the three MRCP(UK) parts.**

<i>Measure</i>	<i>Part I</i>	<i>Part II</i>	<i>PACES</i>
Coefficient Estimate with 95%CI	0.24 95% CI [0.05, 0.44]	0.22 95% CI [0.08, 0.34]	0.22 95% CI [0.12, 0.32]
Tau <sup>2</sup>	0.047 (SE=0.033)	0.021 (SE=0.015)	0.0110 (SE=0.009)
I <sup>2</sup>	99.93%	99.76%	99.57%
H <sup>2</sup>	1434.89	424.66	230.09
Cochran's Q	Q(4)=10,669.88, p<0.001	Q(4)=2,677.42, p<0.001	Q(4)=925.33, p<0.001

The results of the meta-analyses presented in Table 76 show that the effect of MRCP(UK) parts on underperformance criteria was moderately high; however, the parameters of the models indicated high heterogeneity resulting from the selection of studies. This is understandable considering the number of studies included in these meta-analyses; under normal circumstances, five is too low a number for conducting meta-analyses. However, considering the purposes for which these calculations were performed, where meta-analyses serve as means of estimating the average effect size of each MRCP(UK) part on the criteria, it was considered sufficient.

Figure 30 on the next page contains forest plots depicting the observed effects for all MRCP(UK) parts jointly.

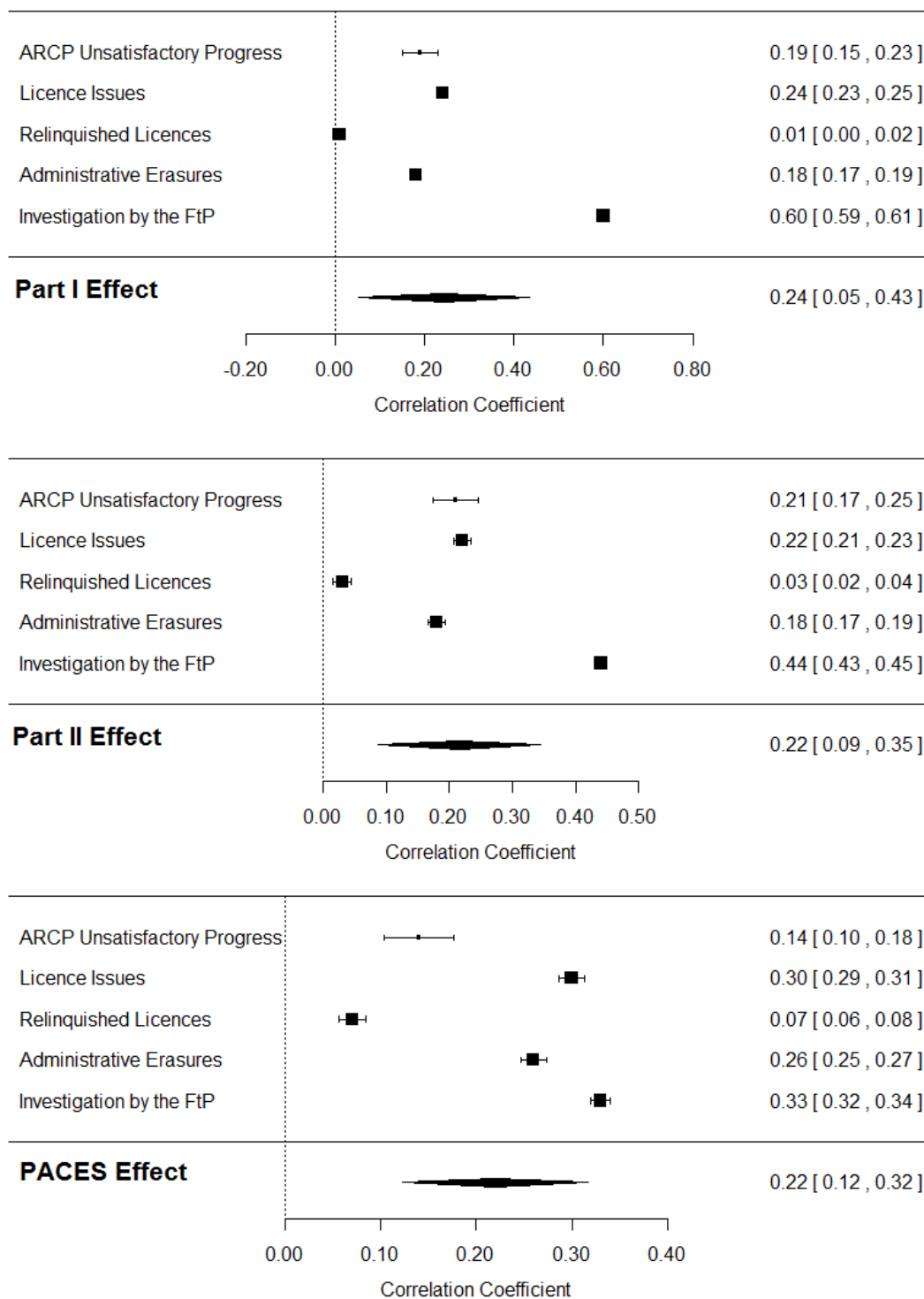


Figure 30. Standard forest plots for effects of MRCP(UK) parts on underperformance criteria.

Despite the low number of cases and high heterogeneity, the funnel plots did not indicate presence of researcher bias in study selection. These funnel plots are provided in Appendix E.



## SUMMARY AND DISCUSSION

The meta-analyses aimed to provide estimates of the average true relationship between each of the MRCP(UK) parts and the criterion measures employed in this research. The models separately estimated these effects based on univariate statistics grouped by the type of investigated criteria: examinations and underperformance measures. In the first case, the models were based on correlation coefficients corrected for artefacts such as attenuation and range restriction, and in the latter case, they were based on mean differences converted into effect sizes (point-biserial correlation coefficients). Random-effects models were employed. The models were fitted under an artificial assumption that Part I, Part II, and PACES coefficients originate from separate samples, as the purpose of the presented meta-analyses was the estimation of the separate effect of each MRCP(UK) part on the criteria.

Random-effects models fitted for correlation coefficients showed a positive moderate to high association between Part I, Part II, and PACES scores and the exams chosen as criterion measures for this research. The estimates of the true effects reached 0.69 for Part I, 0.70 for Part II, and 0.48 for PACES. These results mean that the higher the scores in the MRCP(UK) part, the higher the scores in any of the subsequent assessments. In particular, these results confirmed the conclusion of Chapter 5 that Part II and Part I are the strongest correlates, with Part II being only slightly stronger or effectively equal. This also means that the criterial examinations were likely to be based on both presenting factual knowledge and applying it. The effects of both Parts I and II were also stronger than the effect of PACES (clinical skills).

Inspection of the forest plots showed that MRCGP exams carried the largest weight in the estimates of the mean result, which was due to the large sample size. Other exams were of similar weight and individually had smaller impact. The widest confidence intervals were observed for exams with small sample sizes, e.g. Palliative Medicine, meaning that the estimates of correlations for these exams were provided with lesser precision. Importantly, none of these confidence intervals included zero.

The meta-analyses of underperformance measures, namely Licence Issues, Administrative Erasures, Voluntary Erasures, Underperformance in Training, and GMC FtP investigations, were less robust due to a small number of criteria (five) being incorporated into the models. The estimated effect of MRCP(UK) parts on underperformance criteria reached 0.22 to 0.24 (in absolute values). These estimates were much lower than in the case of

examinations, but it should be remembered that the effect sizes entered into the models were neither disattenuated nor range derestricted. Further, clinical underperformance is a difficult measure to quantify, and in case of this research binary variables were used, which provided less variability in the initial analyses. The results suggest that there was a significant, albeit varied, effect of MRCP(UK) parts on clinical underperformance, which on average should be considered moderate. However, the mere existence of this relationship indicates the relevance of MRCP(UK) examinations to real life clinical situations. The results indicate that candidates who do better at MRCP(UK) also do better in training and are less likely to experience issues with registration.

A general observation from the above meta-analytical models was that the models were highly heterogeneous. However, based on the values of the  $I^2$  statistic, this heterogeneity originated in the studies rather than from sampling error; it resulted from the variability in the influence of MRCP(UK) parts on criterion measures. In the case of the models quantifying the effect on examinations, a plausible explanation could be that due to extreme specialization in medicine, the types of knowledge tested in different exams overlap only to certain extent. Therefore, the medical knowledge itself is not a homogenous construct. This would create an actual difference in how much MRCP(UK) parts can predict or correlate with the specialized exams which were chosen to be the criteria in this research. Other factors, such as the form of the test, its length, or the time interval between the exams, may have potentially added to the heterogeneity. Finally, the observed heterogeneity may have also been the result of the innate differences between the groups of doctors choosing different specialties. Perhaps there are inter-individual factors that make doctors choose, for example, a GP specialty over Clinical Oncology or Renal Medicine. Such differences could also add to the span between the values of the estimated effects. In the case of analyses on underperformance criteria, the heterogeneity resulted from similar reasons, such as internal differences and variability in the construct, but the extent of heterogeneity was likely to result mainly from the small number of analysed studies.

Perhaps a weakness of the presented meta-analyses was that they were not performed on effect sizes grouped by constructs, despite the fact that previous analyses followed this pattern. This was due to the fact that some effects could be corrected for range restriction and disattenuated while others could not, and models comprising both types would be misleading.

In summary, the results of the meta-analyses summarised the effects of MRCP(UK) parts on doctors' subsequent performance in knowledge examinations and clinical skills assessments and on-the-job performance. The strength of the estimates of these effect sizes suggest that MRCP(UK) parts are good predictors of future performance in general, which supports the notion of predictive validity of MRCP(UK).

## Chapter 8

### Discussion of the results and summary

#### ABSTRACT

*The previous chapters presented the background, methodology, and results of the analyses on the relationship between MRCP(UK) parts and measures of clinical performance and knowledge assessments. It was hypothesized that MRCP(UK) would predict the results of such measures, and this hypothesis was confirmed both through univariate and multivariate statistical methods. It was found that in the case of subsequent medical examinations Part II and Part I scores were the best predictors. In the case of MRCGP clinical assessment ('CSA') the best predictor was PACES. In the case of on-the-job performance, MRCP(UK) parts were indicative of disciplinary actions, voluntary erasures, administrative erasures, and performance during specialty training. The magnitude of the coefficients was within the expected range and was concordant with the examples from the literature. The criticisms this research may encounter and directions for future research are addressed. This chapter also presents the place of the above-mentioned results within the body of knowledge on high-stakes medical examinations and discusses the practical implications of the findings.*

The purpose of this PhD thesis on the predictive validity of MRCP(UK) was to investigate whether MRCP(UK) scores relate to subsequent professional performance. It was hypothesised that MRCP(UK), as an exam testing knowledge, skills, and professional behaviours – the very features required of a medical professional – would predict certain criteria chosen as representative of the above-mentioned characteristics. Following the examples from the literature on other high-stakes medical exams, evidence of the predictive validity of MRCP(UK) was sought in the relationships between MRCP(UK) scores and scores in subsequent medical knowledge exams, outcomes of subsequent skills assessments, and in measures of general performance at work. Specifically, it was hypothesised that doctors who perform better in MRCP(UK) parts would also perform better in subsequent examinations and assessments. In addition, it was hypothesised that they would be less likely to experience issues related to underperformance, such as investigations or licence limitations. The results of the performed analyses supported the hypotheses, and findings can be grouped into major clusters: on the relationship between the MRCP(UK) parts, on the relationship between MRCP(UK) and knowledge exams, and

the relationship between MRCP(UK) and clinical assessments. Meta-analyses summarised these results.

## **8.1 SUMMARY OF THE RESULTS**

### **8.1.1 Relationship between MRCP(UK) parts**

The first group of evidence in favour of the predictive validity of MRCP(UK) comprises the relationships between MRCP(UK) parts themselves. Analyses showed that the first attempt scores in Part I predicted both Part II and PACES first attempt scores, and that Part II scores predicted PACES scores. It was also found that the first attempt scores in an MRCP(UK) part are predictive of all consecutive attempt scores in that part, which was already described in McManus and Ludka (2012). The uncorrected correlation coefficients reached 0.60 between first attempt scores at Part I and Part II, 0.30 between first attempt scores at Part I and PACES, and 0.38 between first attempt scores at Part II and PACES. Corrections for range restriction and attenuation increased the coefficients to reach 0.78, 0.43, and 0.48, respectively. The pattern of the observed coefficients indicated congruence of the forms between Part I and Part II and their contents, and incongruence between written Parts I and II and PACES; the coefficients between written exams were stronger than coefficients observed between a written exam and a clinical skills assessment.

### **8.1.2 Relationship between MRCP(UK) and knowledge exams**

The second group of findings supporting the predictive validity of MRCP(UK) related to medical knowledge, indicated previously as one of the key components of medical professionalism. The criteria representing knowledge that were chosen for comparison were based on performance in fifteen medical exams taken post-MRCP(UK), and included: twelve Specialty Certificate Exams ('SCEs'), two SCE equivalent exams (FRCR and CKBA), and the MRCPG AKT examination. The univariate analyses showed that MRCP(UK) scores correlated with all other medical exams scores. The uncorrected correlation coefficients varied from 0.29 to 0.66 for Part I, from 0.32 to 0.68 for Part II, and from 0.16 to 0.50 for PACES. The correlation coefficients were corrected for the artefacts: attenuation (unreliability) and range restriction, in order to estimate the true strength of these relationships. The corrections resulted in those coefficients ranging from 0.46 to 0.86 in the case of Part I, from 0.51 to 0.86 in the case of Part II, and from 0.21 to 0.68 in the case of PACES. The differences in the strength of the coefficients showed a pattern consistent with the notions of psychometric theory (Anastasi & Urbina, 1997; Cronbach, 1970). Based on the literature on validity, the strength of validity coefficients is dependent on the similarity

of the constructs measured by the tests, their form and other common variance effects, the time interval between them, and their unreliability. Of these factors only three could potentially explain the observed differences, as the issue of unreliability was addressed during the process of correcting for range restriction. Apart from the desired effect of congruency between the constructs, other factors could still have had an impact on the magnitude of the correlation coefficients. With respect to the form of the exams, indeed the coefficients were higher when two written exams were correlated in comparison to a written exam and a clinical assessment. This was observed for MRCP(UK) parts, as already mentioned, but also for all other exams: coefficients associated with Part I, Part II and SCEs were higher than for PACES and SCEs. The direct effect of the time-interval on coefficients could not have been confirmed with absolute certainty; however, certain observations may imply the existence of the effect of time. The coefficients associated with MRCGP AKT were comparable or slightly higher than the majority of those associated with SCEs, and the SCEs coefficients were generally higher than those associated with the FRCR exams. These differences coincide with the mean time intervals, as MRCGP is on average attempted 1 to 3 years after MRCP(UK), while specialty certificate exams may be attempted 4 to 6 years later, with FRCR examinations being taken after the longest period of time. The obtained correlation coefficients could have also been affected by the common method variance; however, the multitude and independence of sources of data and variety of methods with which the data were collected should significantly decrease the probability that the observed relationships are spurious. Yet, a separate investigation would be required to separate the effects of the confounding factors influencing the magnitude of the coefficients.

The results of the univariate statistics were confirmed by the multivariate analyses, which regressed MRCP(UK) parts scores onto scores in knowledge assessments. The purpose of these analyses was to quantify the joint effect of MRCP(UK) parts on the criterion measures. The analyses showed that the best predictors were either Part II or Part I scores. Nine out of sixteen fitted models excluded PACES scores as non-predictive. The models where PACES was a significant predictor were fitted for MRCGP AKT, Acute Medicine, Endocrinology, Gastroenterology, Geriatric Medicine, Rheumatology, Cardiology, and FRCR1. However, in each of these models PACES scores had a much smaller impact on the dependent variable than the written Parts I and II. Despite the fact that the fitted regression models varied in the number of significant predictors and associated coefficients values, comparison between them – performed either with multi-level modelling or Chow

tests – showed no significant difference; the only exception was the MRCGP AKT model. This indicated that the fifteen fitted models, statistically speaking, could be considered parallel in terms of explaining the relationship between MRCP(UK) and the criteria chosen to represent knowledge. The model for the oral component of the FRCR2 examination was altogether non-significant, and therefore was excluded from testing for similarity. These results demonstrate a strong relationship between MRCP(UK) and the knowledge exams, particularly in the ability to apply knowledge in interpreting medical data, which confirms the validity of MRCP(UK) as an exam testing this particular aspect of medical professionalism.

### **8.1.3 Relationship between MRCP(UK) and clinical skills and performance measures**

The third group of evidence describes the relationship between MRCP(UK) and the criteria based on clinical skills assessments and on-the-job performance. These criteria were assumed to represent jointly the second and third pillars of professionalism, which are clinical skills and professional attitudes. These two aspects were investigated jointly, as substantiated in detail previously in Chapter 2. Briefly summarising, it was impossible to separate the expression of appropriate attitudes from possessing adequate clinical skills at the measurement or criterion level. The standardised clinical assessments and on-the-job performance assessments usually amalgamate these two aspects into one joint outcome. The criteria within this group of evidence were two exams (the MRCGP CSA and FRCR2 clinical exam), the ARCP outcomes as a general assessment of performance during training, general underperformance based on registration status, and also the status of being investigated by the GMC FtP panel. The LRMP statuses that were taken into account in the analyses included the Licence Issues as a collective measure of licence limitations or a disciplinary erasure, Voluntary Erasure (as sometimes it is granted in the course of investigation process by the Fitness to Practice panel in order to avoid disciplinary erasure), and the Administrative Erasure as a distant proxy for conscientiousness, and the only criterion that was considered an entirely separate measure of attitudes. The results of the analyses with respect to the above-mentioned measures generally supported the hypothesis for this research. The three parts of MRCP(UK) correlated with the MRCGP CSA (one of the clinical examinations), and the coefficient associated with PACES was the highest ( $r=0.46$  uncorrected;  $0.59$  corrected). The only significant correlation found between FRCR2 clinical component and MRCP(UK) was that with Part II scores (uncorrected  $r=0.21$ ); no significant correlation was found with PACES or Part I scores. This suggests that CSA and the FRCR2 clinical exam differ in terms of what they measure. Constructs behind

CSA and PACES are more closely related than in the case of PACES and FRCR2. Whilst speculative, it is possible that FRCR2 is highly specialised, which could explain the lack of correlation. Another plausible explanation is that the number of cases on which the analyses were performed was too small to capture the effect.

The relationship between MRCP(UK) performance and unsatisfactory progress in training (measured with standard ARCP outcomes) was established via a comparison of candidates' mean scores at their first attempt at Part I, Part II, and PACES. The comparison was carried out between those who had all satisfactory ARCP outcomes throughout their training, and those who had at least one unsatisfactory outcome. The differences between the two groups were in each case statistically significant and always favoured trainees with satisfactory progress (effect sizes:  $r=0.19$  for Part I,  $r=0.21$  for Part II, and  $r=0.14$  for PACES scores).

Statistically significant differences were also revealed when the MRCP(UK) scores of underperforming doctors were compared with scores obtained by their colleagues in good standing, i.e. not experiencing any issues with registration in the time frame of four years for which the LRMP data were available. The investigated differences were based on the registration status and GMC Fitness to Practice reviews. A comparison was made of mean first attempt scores in the three parts of the MRCP(UK) between those who experienced Licence Issues and those who did not. This revealed that doctors from the former group had significantly lower scores in MRCP(UK) than their colleagues in the latter group. The effect sizes were moderate, ranging from  $r=0.22$  to  $r=0.30$ . These results were corroborated by the analyses performed on doctors who were on the GMC FtP panel list and had a score in the MRCP(UK) History File, even though only eight such cases were identified. Comparison of mean scores between these doctors who were investigated by the GMC and the rest of the doctors in the History File showed that the GMC FtP listed medics scored on average 14% lower in Part I ( $r=0.60$ ), 6.2% lower in Part II ( $r=0.44$ ), and also scored almost 5% lower in PACES ( $p=0.33$ ). Considering the number of cases these effect sizes are impressive, and what is worth noting, much higher than those observed for Licence Issues. Doctors identified on the GMC FtP panel list were going through a full review, including examination, while doctors who experienced Licence Issues more often had milder forms of disciplinary actions imposed on them, e.g. undertakings. The differences between these effect sizes may be interpreted as revealing a connection between examinations performance and the severity of the disciplinary actions.



Disciplinary erasure from the register yielded a higher effect size than Voluntary Erasures, which are on occasion granted to doctors who go through the investigation process. Statistical analyses of differences between those who voluntarily erased themselves from the register and the rest of the doctors on the register did not show significant differences apart from PACES. The effect size was, however, extremely small ( $r=0.07$ ), suggesting that reasons for relinquishing a licence are in general unlikely to be associated with performance. Investigation into differences between doctors who had their registration erased for administrative reasons and the rest of the doctors yielded significant results. Doctors who were erased scored significantly lower in all MRCP(UK) parts than their registered colleagues, and the effect sizes associated with these differences were moderately high ( $r=0.18$  for Part I and Part II,  $r=0.26$  for PACES). Considering that Administrative Erasures were only a far proxy of conscientiousness, the effects are surprisingly large. A possible explanation could be that Administrative Erasures are (albeit not explicitly) driven by factors other than conscientiousness alone. Arguably, it is possible that the group of administratively erased doctors comprise large numbers of international doctors who, for example, decided not to practice medicine in the UK, and instead of applying for relinquishment simply did not proceed with necessary administrative duties. This seems to be a plausible explanation judging from the odds ratios associated with ethnicity and PMQ obtained through multivariate analyses.

The multivariate analyses on underperformance were employed to establish the joint effect of MRCP(UK) parts on the criteria. For each of the criterion measures a regression model was fitted. These models showed that PACES was the best predictor for MRCPG CSA, with Part I and Part II still significant, but with coefficients of approximately half the size. The model for the FRCP2 clinical exam was non-significant, which was most likely due to the low number of valid cases; however, it did indicate an impact of PACES and Part II scores on the FRCP2 scores. A logistic regression model for ARCP showed that the best predictors were lower Part II scores and lower Part I scores, with PACES being non-significant. Logistic regression models for registration statuses corroborated the influence of MRCP(UK) parts seen in univariate analyses, even after taking into account the demographic factors. For example, in the case of Licence Issues, male sex, older age, and lower PACES scores were all significant predictors in the model. Voluntary erasure was predicted by higher PACES results, having a non-UK primary medical qualification, and being older in age. Further, higher Part I results and lower PACES results were independent predictors of being erased for administrative reasons, while being a UK graduate, being white, and being younger all

decreased the likelihood of being removed from the LRMP based on an administrative decision. The significant effect of demographic factors such as ethnicity, PMQ, and gender on performance found in the course of this research was consistent with the findings of other researchers (Dewhurst *et al.*, 2007; Haq *et al.*, 2005; Humphrey, Hickman, & Gulliford, 2011; McManus & Wakeford, 2014; Tiffin *et al.*, 2014; Woolf *et al.*, 2011); however, no clear indication to the reasons for the presence of these effects has been available. Possible explanations included differences in training programs (McManus & Wakeford, 2014), language proficiency, or learning styles and personality factors (Haq *et al.*, 2005); however, further research is required.

#### **8.1.4 Results of the meta-analyses**

In order to summarize the above-mentioned findings and to estimate the size of the average effect of each of the MRCP(UK) parts on the examination criterion measures and underperformance measures, six meta-analytical models were fitted. The first group of models performed on examinations (on knowledge tests and standardised clinical skills assessments jointly) was based on correlation coefficients corrected for range restriction and disattenuated. The model indicated that the average correlation coefficient between Part I and the criteria reached 0.69 with 95%CI [0.64, 0.75]. In the case of Part II the average effect was estimated at 0.70 with 95%CI [0.66, 0.74], and in the case of PACES the effect was estimated at 0.48 with 95%CI [0.42, 0.53]. These effects should be considered high, which confirms that MRCP(UK) scores are strongly tied to performance in subsequent examinations. The models for clinical underperformance were less robust and yielded smaller absolute estimates of  $r=0.22$  to  $0.24$ . The higher estimate was associated with PACES; however, the differences in estimated true effect sizes are extremely small. The average effect sizes were most likely fuelled by a very strong effect observed for the GMC FtP panel investigations ( $r=0.33$  to  $0.60$ ), and lowered by incorporation of Voluntary Erasures into the meta-analytical models. Should there be a larger number of coefficients to meta-analyse, neither of the two effects would have such an influence on the final results.

All of the above groups of evidence consistently indicate that MRCP(UK) scores are predictive of performance in subsequent tests, clinical skills assessments, and on-the-job performance. Taken together this supports the hypothesis of the predictive validity of MRCP(UK). Further, the strength of the obtained validity coefficients was consistent not only with the assumptions based on theoretical notions of psychometrics, but also concurred with the literature on medical exams. For example, in a study by Simon *et al.* the correlation coefficient between OSCE and USMLE1 reached 0.41 (Simon, Volkan, Hamann,

Duffey, & Fletcher, 2002). The coefficients between USMLE and the in-training exams for the certification of the American Board of Paediatrics ranged from 0.65 for Step 1 to 0.79 for Step 2 (McCaskill *et al.*, 2007). The same study has also shown that USMLE exam scores were correlated with the final certification exam with  $r=0.67$  and  $0.63$ , respectively. In a study focused on predictive validity of MCAT, the coefficient between MCAT and USMLE1 reached  $0.61$  (Julian, 2005), which was confirmed later by Donnon *et al.* (2007), where the weighted effect size for the same pair of tests was  $0.60$ . In the latter study, the influence of MCAT on basic science/preclinical test results in medical schools was averaged to  $r=0.39$ . A similar correlation level was found between Flemish Admission Test scores and the final first year medical school score; it reached  $0.35$  (Lievens & Coetsier, 2002). Also, smaller correlation coefficients were associated with more specific tests, such as verbal reasoning subtests ( $r=0.19$ ) or physical sciences subtests ( $r=0.23$ ) of MCAT (Donnon *et al.*, 2007). Within the UK medical education system, a study by Ahmed, Rhydderch, and Matthews (2012) showed that the selection procedure for the general practice training is predictive of future AKT and CSA results, with coefficients  $r=0.49$  and  $r=0.53$  respectively. Also, in terms of professional behaviour, the literature provides values of coefficients that can be compared with the results obtained within this study. For example, in a study by Stern *et al.* (2005) a multilevel regression model indicated that certain behaviours signifying a lack of conscientiousness were predictive of subsequent unprofessional behaviour (as found by a review board). The beta coefficients were estimated at  $\beta=0.23$  for lack of course evaluations and  $\beta=0.29$  for failing to report immunisation compliance. A study by Kelly, *et al.* (2012) showed that the Clinical Conscientiousness Index results correlated with professionalism as perceived by the faculty members at  $r=0.30$  and with the OSCE performance ( $r=0.24$ ). It is also highly relevant that the referred-above examples from the literature showed a similar effect of congruency between the constructs that was observed in the course of this research. The coefficients between two written knowledge exams were found to vary around  $0.60$ , while the reported coefficient between clinical skills assessments and written test reached  $0.40$ .

## 8.2 LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

As previously mentioned, this research was designed with the acknowledgement of certain limitations. Firstly, it was designed as purely quantitative study. However, within the course of this project it became apparent that a meaningful interpretation of some of the obtained results, such as the decision on voluntary erasure, demands an insight from a qualitative study that would provide a context for individual decisions. Therefore, one way of

continuing this research would be to design a qualitative study or a quantitative one with a qualitative component that would provide more depth to the inferences made based on statistical evaluation of dichotomised measures.

Secondly, this project was designed based on retrospective longitudinal data limited to eight years – the main dataset of candidates contained records from May 2003 to January 2011. Although this timeframe provided more than 50,000 records of MRCP(UK) candidates for analyses, the actual effective number was often significantly smaller. This was partially due to the limited sample sizes of criterion measures datasets. Apart from LRMP and MRCGP, which provided a large number of records, other examinations and assessments were taken by smaller numbers of doctors. Secondly, a certain number of doctors within each dataset did not have a GMC number, which was the primary matching key. This was mostly an issue with the records of International Medical Graduates. As a result the smallest matched sample was 8 doctors for the GMC list of doctors under investigation, and the second smallest was 31 for the Palliative Medicine specialty certificate examination. Thirdly, the numbers of valid cases were also uneven when considering each of the datasets and a different MRCP(UK) part, which was due to the censoring that was present in the data. Nonetheless, the median sample size was still approximately 570 cases, which should be sufficient considering the reduction of sampling error with increasing sample sizes.

The retrospective longitudinal design also effectively limited the pool of suitable criterion measures, as they could have only been sourced from pre-existing data rather than being purposefully designed. This may have resulted in clinical criteria being underrepresented in comparison to knowledge exams, which were relatively abundant and easy to secure. The issue of underrepresentation of clinical criteria that affected this research, however, is a larger problem. It stems from the foundations of the system of ensuring medical quality in the UK. Details of cases of medical errors, individual indices of contraindicative prescribing, or individual ratios of prescribed screening tests are simply not collected in the UK, whereas they are in the US and Canada. For example, in a recent study Norcini *et al.* (2014) were able to link the mortality rates in the US with the performance of international doctors in USMLE Step 2 examinations. Such a study would not be possible in the UK setting due to the lack of appropriate measures for monitoring doctor' individual performance. As already mentioned in the introductory Chapter 2 (section 2.2.2), attempts of introducing tighter monitoring met with counterarguments that quantified individual measures may lead to misinterpretation (Lanier *et al.*, 2003). However, it seems these fears are unsubstantiated. Current statistical methods and computational power allows for distinguishing the effects

of lack of competence from severity of cases, as in Norcini *et al.*'s (2014) study. Being able to connect the medical exam results with quantifiable clinical outcomes is the 'holy grail' of predictive validity, and further research focusing on this aspect of any UK medical exam will have to face the situation of insufficient direct measures, as was in the case of the current research. Future studies should investigate the predictive validity of MRCP(UK) after securing data that provide measures of clinical errors. Importantly, they should also include measures of attitudes; these were scarce when the data collection process for this research was taking place. One method of collecting such data on attitudes, in order to quantify them and to test if MRCP(UK) can measure and predict them, could be a questionnaire purposefully designed for current students and Foundation Years doctors. Such a tool could be similar to the Conscientiousness Index proposed by McLachlan *et al.* (2009), and if applied on larger groups of doctors annually over a longer period of time, it could provide insight into the level of the professional behaviours and attitudes, and their development.

Another aspect of this research that could meet with potential criticism is that it should have been able to equally examine those candidates who passed MRCP(UK) and those who failed it. However, the majority of the available suitable criteria were the knowledge exams, which are only taken after passing MRCP(UK). This effect (known as range restriction) is pertinent to all entry or qualifying examinations that can be failed, and it lowers the variability of the scores and validity coefficients. Although derestriction of range was applied whenever possible, the corrected coefficients are the estimates of the true strength of the relationship between MRCP(UK) and the criteria, which inherently assumes a certain level of error. The only two criterion sources that had information on candidates who failed MRCP(UK) were the LRMP and, to a certain extent, MRCGP. Therefore, it might be argued that this study focused mainly on those candidates who were successful in MRCP(UK). Looking into candidates who failed MRCP(UK) and their career paths might be yet another direction future studies could pursue.

This research was also limited as it focused on finding general relationships between MRCP(UK) parts overall scores and the available criterion measures. However, MRCP(UK) written parts consist of blocks of questions in several fields of medicine, while PACES scores are based on generic medical skills. In both cases such sub-scores could potentially be incorporated into the analyses and correlated with the criteria. This would be another possible avenue of research to extend the findings of the current study. In the case of PACES, this particular direction could be taken within two to three years, when candidates who passed the new PACES go through the ARCP assessments and attempt SCEs. At the

time of conducting this research, there was insufficient time since the introduction of the new PACES scoring (in 2009) to connect the candidates' scores with subsequent clinical assessments or on-the-job performance.

Apart from investigating the above-mentioned aspects, future studies could also focus on extending the body of work presented in this thesis by increasing the time-frame of both the criterion and MRCP(UK) datasets to obtain more stable results from larger and wider samples. In view of the lack of hard quantifiable data on clinical performance it would be advisable to look into individual cases of misconduct and malpractice and classify those cases into categories of contraindicative prescribing, medical errors, and lack of diligence, as done in Canada and the US, to investigate the relationship between such measures and the MRCP(UK) performance. Effectively, this would mean expanding the current list of sources of criterion measures. Additionally, securing access to the results of candidates who have attempted MRCP(UK) and subsequently decided to take another career path in another college that did not provide data for this research would be advisable. This may also mean looking into the careers of those who ultimately failed MRCP(UK) and decided not to practice medicine at all.

### **8.3 MEANING OF THE FINDINGS AND GENERAL IMPLICATIONS**

All of the above limitations do not, however, diminish the meaning of the presented findings. This research is the first that holistically addresses the issue of predictive validity of a major postgraduate examination in the UK. Previous research has either addressed admissions exams or admissions procedures (e.g. Ahmed *et al.*, 2012; Emery & Bell, 2009; McManus *et al.*, 2011; McManus, Dewberry, *et al.*, 2013), or related to the general validity of an exam (e.g. Metcalfe, 2012). The abundance of evidence on the relationships between MRCP(UK) and criteria chosen for this study, in particular their strength and consistency, show that there is an underlying factor or factors that fuel the statistically significant interdependence. Since these factors manifest themselves through performance in medical exams, medical assessments, and doctor's on-the-job performance, it can be assumed that they comprise the requisite knowledge and skills to be a medical practitioner. Therefore, it can be assumed that the unifying factor for all relationships observed over the course of this research is medical competence, which constitutes a strong argument in favour of the predictive validity of MRCP(UK). At the same time it provides arguments for the validity of the criterial exams and assessments.

On the other hand, success in an examination may not only be driven by knowledge or skills, but can also be aided by personal characteristics, such as aptitude, conscientiousness, and the ability to handle stress, as well as having a supportive lifestyle, and an absence of negative life events. Hence, it may be argued that the results presented here do not support the quality of MRCP(UK) design, but rather indicate that such personal aspects are predominant factors in passing a test. Indeed, the effect of personal values on success in e.g. medical school has been investigated before, and according to the literature conscientiousness and learning styles, among other factors, do explain variability in the performance to certain extent (Ferguson *et al.*, 2002) even when taking previous academic achievements into account. This research acknowledged that personal features constitute a factor affecting performance by taking into account the proxy measure of conscientiousness as a criterion. Possible arguments regarding the candidates' aptitude being the dominant predictor of success were already addressed in Chapter 5, where the concept of the Academic Backbone and the predictive validity of aptitude tests in medical student selection were discussed (McManus, Woolf, *et al.*, 2013; McManus, Dewberry, *et al.*, 2013). According to the studies referenced here, aptitude itself is not a sufficient factor to succeed in medical school, but rather constitutes a foundation upon which knowledge can accumulate. Should aptitude be the only responsible factor for success in an exam, the correlation coefficients between two exams should be of similar strength irrespective of the field of testing. The findings of this research show that the strength of the validity coefficients varied across knowledge tests depending on their contents, which was particularly visible in the case of MRCGP AKT sub-scales scores, making the hypothesis that aptitude takes a predominant role questionable. However, the direct effect of aptitude was not analysed in the course of this research.

It may further be argued that despite the observed associations between MRCP(UK) and the criterion measures, these relationships are not of predictive nature, as this thesis did not provide indication that certain results would inherently lead to underperformance in medical practice. Indeed, the majority of the analyses are correlational which *per se* does not imply causality. However, it was a prerequisite of this study that all criterion sources were found to occur post-MRCP(UK) and were embedded in medicine. In the field of medical education it was shown through the Academic Backbone (McManus, Woolf, *et al.*, 2013) that success at consecutive stages of the educational process requires continuous accumulation of knowledge ('cognitive capital'), becoming the causal link between each stage, either explicitly (i.e. when an event A causes event B) or at a construct level (i.e.

where events A and B are both driven by C). Therefore, it seems unlikely that correlations obtained and presented here were spurious. Further, on a more detailed level, this research employed regression models, which assume causality by definition. The fact that a particular outcome cannot be ascertained is another issue that results directly from the probabilistic nature of models. Models are a simplified version of the observed world and as such are bound to provide at best probabilistic approximations of reality. In this sense every regression model provided in this thesis was such an approximation: of failing an exam, of failing a clinical skills assessment, of experiencing issues with registration or in training; all based on the MRCP(UK) scores. These models may have fallen short in terms of certain statistical measures, such as e.g.  $R^2$ , or statistical significance, but noise can sometimes obscure the signal (Silver, 2013), and the medical field is particularly susceptible to various factors affecting performance. This raises the question of whether the quality of the models presented in this thesis could have been better. The response would be that, apart from the appropriateness of methods serving its fitting, the quality of a model stems from the amount and the quality of data it is built on. Since the data collected for the purposes of this research were considered reliable, the models should also be considered reliable, and the only solution to make future models better is to acquire more predictors and more data.

Among the presented results, one delivers a particularly important message: the fact that doctors who experienced trouble with their registration and who underperformed, did less well at MRCP(UK). This supports the notion that MRCP(UK) is a valuable selection assessment, and that the standards with which it is administered affect the standards of medical care in the UK. The literature provides several examples where valid major medical examinations were associated with better medical care (for example, Norcini *et al.*, 2000, 2002; Reid *et al.*, 2011; Sharp *et al.*, 2002; Tamblyn *et al.*, 1998, 2002). Therefore, the predictive validity of MRCP(UK) is by extension an affirmation of competence and professionalism. This is fundamental for the public to trust the medical profession, even if the observed effects are moderate in magnitude.

The presented findings on predictive validity vouch for the quality of MRCP(UK) and justify maintaining it as a selection process, by providing a long-term insight into the medical careers and consequences of its use. As approximately 6,300 doctors attempt MRCP(UK) annually, the findings of this research can provide them with answers as to why MRCP(UK) is an important part of their medical career path, should they wish to practice hospital medicine in the UK. It should provide encouragement that their efforts in passing this exam



will reflect in their future good performance. At the same time, the public will find arguments that the quality of MRCP(UK) warrants that only appropriate prospective candidates are chosen, with the intention of the medical profession to ensure the best possible medical care. Therefore, the evidence on the predictive validity of MRCP(UK) supports the notion that its administration translates into a better quality of medical services, and decreased rates of medical errors.

## **SUMMARY**

The results of this thesis have supported the hypothesis that MRCP(UK) is a valid predictor of subsequent performance in examination, clinical skills assessments, and on-the-job performance. The MRCP(UK) parts scores predicted the results of the knowledge tests and the clinical performance measures that were part of this research. The correlation coefficients between MRCP(UK) parts and the criteria were moderately high or high, as expected based on the literature. The coefficients associated with medical knowledge assessments showed the expected effect of congruency of the constructs: the relationship between written parts of MRCP(UK) and criterion written tests was stronger than between PACES and the written tests. The clinical skills assessments showed no consistent pattern; however, PACES was the best predictor in the case of CSA. The fact that PACES was a non-significant predictor in the case of clinical FRCR2 examination can be explained by the low number of valid cases, and rather substantial differences in the contents of the two exams. This issue could be potentially investigated further when more data are available.

The analyses also showed that doctors who had any difficulty during specialty training scored lower in all three parts of MRCP(UK); however, the regression model for ARCP indicated that Part II and Part I had a predominant role in predicting the unsatisfactory outcome. One important piece of evidence in favour of the predictive validity of MRCP(UK) came from its parts being predictive of general underperformance quantified as: Licence Issues Relinquished Licences and Administrative Erasures, and being investigated by the Fitness to Practice panel. Firstly, doctors investigated by the GMC FtP panel scored lower in the three parts of MRCP(UK), with quite large differences in Part I results. Secondly, a relationship was found between MRCP(UK) and Licence Issues and Administrative Erasures. Those who experienced any of the two scored lower in all three parts of MRCP(UK). A logistic regression model indicated that a low PACES score had a predominant role in the experiencing of Licence Issues, together with age and being male. Removal from the

register for administrative reasons was also associated with lower PACES results (and higher Part I results) together with other demographic factors.

The results presented in this thesis add to the existing body of knowledge on high-stakes medical examinations, and indicate the need to investigate the predictive validity of such exams through linking the scores with external criteria over a period of time. As discussed, such suitable criteria should preferably be found among the measures of actual clinical performance, as is the case for studies from Canada and the US. Currently, such measures are unavailable in the UK setting which was a major limitation in the case of this research, and which will present an obstacle for any future predictive validity study of a UK examination. However, irrespective of the criticisms addressed above, in the light of the presented evidence, MRCP(UK) should be considered a valid exam predictive of subsequent assessments and general clinical performance. These results carry particular weight for both doctors who take the MRCP(UK) and for the public, as they indicate that MRCP(UK) is a valid tool for de-selecting candidates who may underperform in their medical practice, which translates into higher quality and safety of medical services.

*Prediction is difficult for us for the same reason that it is so important: it is where objective and subjective reality intersect. Distinguishing the signal from the noise requires both scientific knowledge and self-knowledge: the serenity to accept the things we cannot predict, the courage to predict the things we can, and the wisdom to know the difference.*

*Nate Silver, The Signal and the Noise*

## Bibliography

- AERA, APA, & NCME. (2004). Standards for educational and psychological testing. Washington DC: American Educational Research Association.
- Ahmed, H., Rhydderch, M., & Matthews, P. (2012). Do general practice selection scores predict success at MRCGP? An exploratory study. *Education for Primary Care: An Official Publication of the Association of Course Organisers, National Association of GP Tutors, World Organisation of Family Doctors*, 23(2), 95–100. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/22449464>
- Allsop, J. (2006). Regaining Trust in Medicine: Professional and State Strategies. *Current Sociology*, 54(4), 621–636. doi:10.1177/0011392106065093
- Al-Mahroos, F. (2009). Construct validity and generalizability of pediatrics clerkship evaluation at a problem-based medical school, Bahrain. *Evaluation & the Health Professions*, 32(2), 165–83. doi:10.1177/0163278709333149
- Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). Upper Saddle River, New Jersey: Prentice-Hall Inc.
- Archer, J., Regan de Bere, S., Bryce, M., & Nunn, S. (2014). *Understanding the rise in Fitness to Practise complaints from members of the public*. [plymouth.ac.uk](http://www5.plymouth.ac.uk/uploads/production/document/path/0/75/Archer_et_al_FTP_Final_Report_30.01.2014.pdf). Plymouth. Retrieved from [http://www5.plymouth.ac.uk/uploads/production/document/path/0/75/Archer\\_et\\_al\\_FTP\\_Final\\_Report\\_30.01.2014.pdf](http://www5.plymouth.ac.uk/uploads/production/document/path/0/75/Archer_et_al_FTP_Final_Report_30.01.2014.pdf)
- Arnold, L., & Stern, D. T. (2006). What is Medical Professionalism? In *Measuring Medical Professionalism* (pp. 15–28). Oxford: Oxford University Press.
- Bacon, L., Lynd Bacon & Associates Ltd, & SPSS Inc. (1997). *Using Amos for structural equation modeling in market research*. SPSS. Retrieved from <http://www.bauer.uh.edu/jhess/documents/3.pdf>
- Bagozzi, R., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, 16(1), 74–94. Retrieved from <http://link.springer.com/article/10.1007/BF02723327>

- Baron, J., & Norman, M. (1992). SATs Achievement Tests, and High-School Class Rank as Predictors of College Performance. *Educational and Psychological ...*, 52, 1047–1056. Retrieved from <http://www.psych.upenn.edu/~norman/SAT.pdf>
- Barrett, G. V., & Depinet, R. L. (1991). A reconsideration of testing for competence rather than for intelligence. *The American Psychologist*, 46(10), 1012–24. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/1746769>
- Batenburg, V., Smal, J. A., Lodder, A., & Melker, R. A. De. (1999). Are professional attitudes related to gender and medical specialty ? *Medical Education*, 33, 489–492.
- BCS. (2010). Standard Operating Procedure for the British Cardiovascular Society's Knowledge Based Assessment. Retrieved from [http://www.bcs.com/documents/FINAL\\_Standard\\_Operating\\_Procedure\\_v6.pdf](http://www.bcs.com/documents/FINAL_Standard_Operating_Procedure_v6.pdf)
- BCS. (2013a). KBA : What it looks like British Cardiovascular Society. Retrieved from [http://www.bcs.com/pages/page\\_box\\_contents.asp?PageID=792](http://www.bcs.com/pages/page_box_contents.asp?PageID=792)
- BCS. (2013b). Knowledge Based Assessment in Cardiology. Retrieved from [http://www.bcs.com/pages/page\\_box\\_contents.asp?PageID=526](http://www.bcs.com/pages/page_box_contents.asp?PageID=526)
- Bessant, R., Bessant, D., Chesser, a, & Coakley, G. (2006). Analysis of predictors of success in the MRCP (UK) PACES examination in candidates attending a revision course. *Postgraduate Medical Journal*, 82(964), 145–9. doi:10.1136/pmj.2005.035998
- Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. *Critical Care (London, England)*, 9(1), 112–8. doi:10.1186/cc3045
- Bielby, W., & Hauser, R. (1977). Structural equation models. *Annual Review of Sociology*, 3, 137–161. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/0470011815.b2a13089/full>
- Billington, L., & Taylor, R. (2008). Public Trust in High-Stakes Assessment and its Measurement. Retrieved from <http://cerp.aqa.org.uk/research-library/public-trust-and-high-stakes-assessment>
- Blendon, R. J., Benson, J. M., & Hero, J. O. (2014). Public Trust in Physicians — U.S. Medicine in International Perspective. *New England Journal of Medicine*, 371, 1570–1572. doi:10.4232/1.11759.
- Boenink, A. D., Oderwald, A. K., De Jonge, P., Van Tilburg, W., & Smal, J. A. (2004). Assessing student reflection in medical practice. The development of an observer-rated instrument: reliability, validity and initial experiences. *Medical Education*, 38(4), 368–77. doi:10.1046/j.1365-2923.2004.01787.x
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons. Retrieved from <http://link.springer.com/article/10.1007/BF02293905>
- Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). Fixed-Effect versus Random-Effects Models. In *Introduction to meta-analysis* (Vol. 58). Wiley and Sons, Ltd. doi:10.1016/j.jcv.2013.09.001

- Boshuizen, H. P. A., & Schmidt, H. (1992). On the role of Biomedical Knowledge in Clinical Reasoning by Experts, Intermediates and Novices. *Cognitive Science*, 16, 153–184. Retrieved from <http://www.sciencedirect.com/science/article/pii/036402139290022M>
- Boulet, J. R., Murray, D., Kras, J., Woodhouse, J., McAllister, J., & Ziv, A. (2003). Reliability and validity of a simulation-based acute care skills assessment for medical students and residents. *Anesthesiology*, 99(6), 1270–80. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/14639138>
- Bourne, R. B., Chesworth, B. M., Davis, A. M., Mahomed, N. N., & Charron, K. D. J. (2010). Patient satisfaction after total knee arthroplasty: who is satisfied and who is not? *Clinical Orthopaedics and Related Research*, 468(1), 57–63. doi:10.1007/s11999-009-1119-9
- Brennan, T., Horwitz, R., Duffy, F. D., Cassel, C., Goode, L., & Lipner, R. S. (2004). The role of physician specialty board certification status in the quality movement. *Journal of the American Medical Association*, 292(9). Retrieved from <http://jama.ama-assn.org/content/292/9/1038.short>
- Buller, M. K., & Buller, D. B. (1987). Physicians' communication style and patient satisfaction. *Journal of Health and Social Behavior*, 28(4), 375–88. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/15359948>
- Burns, R., & Burns, R. (2009). Logistic Regression. In *Business Research Methods and Statistics using SPSS*. Sage Publications Ltd. Retrieved from <http://www.uk.sagepub.com/burns/website material/Chapter 24 - Logistic regression.pdf>
- Burrows, P. J., Bingham, L., & Brailovsky, C. A. (1999). A Modified Contrasting Groups Method Used for Setting the Passmark in a Small Scale Standardised Patient Examination. *Advances in Health Sciences Education : Theory and Practice*, 4(2), 145–154. doi:10.1023/A:1009826701445
- Calman, K. C., Temple, J. G., Naysmith, R., Cairncross, R. G., & Bennett, S. J. (1999). Reforming higher specialist training in the United Kingdom--a step along the continuum of medical education. *Medical Education*, 33(1), 28–33. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10211274>
- Calnan, M. W., & Sanford, E. (2004). Public trust in health care: the system or the doctor? *Quality & Safety in Health Care*, 13, 92–97. doi:10.1136/qshc.2003.009001
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait - Multimethod Matrix. *Psychological Bulletin*, 56, 81–105.
- Chamberlain, J. (2010). Governing Medicine: Medical Autonomy in the United Kingdom and the Restratisation Thesis. *Journal of Medical Humanities & Social Studies of Science and Technology*, 1(3), 1–16. Retrieved from <http://www.ea-journal.com/art1.3/Governing-Medicine.pdf>

- Chaytor, A. T., Spence, J., Armstrong, A., & McLachlan, J. C. (2012). Do students learn to be more conscientious at medical school? *BMC Medical Education*, 12(1), 54. doi:10.1186/1472-6920-12-54
- Chi, M. T. H. (2006). Laboratory Methods for Assessing Experts' and Novices' Knowledge. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge Book of Expertise and Expert Performance* (pp. 167–184). Cambridge University Press.
- Chow, G. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society*, 28(3), 591–605. Retrieved from <http://www.jstor.org/stable/10.2307/1910133>
- Cizek, G. J. (2012). Defining and Distinguishing Validity : Interpretations of Score Meaning and Justifications of Test Use. *Psychological Methods*, (January). doi:10.1037/a0026975
- Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of Validity Evidence for Educational and Psychological Tests: A Follow-Up Study. *Educational and Psychological Measurement*, 70(5), 732–743. doi:10.1177/0013164410379323
- Clark, G. (1965). History of the Royal College of Physicians of London. *British Medical Journal*, (1), 79–82. Retrieved from <http://www.bmj.com/content/1/5427/79.full.pdf>
- Cohen, J. J. (2006). Professionalism in medical education, an American perspective: from evidence to accountability. *Medical Education*, 40(7), 607–17. doi:10.1111/j.1365-2929.2006.02512.x
- Cohen, R. J., & Swerlik, M. E. (2002). *Psychological testing and assessment: an introduction to tests and measurement* (5th ed.). McGraw-Hill Higher Education.
- Cooke, A. M. (1972). *A History of the Royal College of Physicians of London, volume 3*. Oxford: Clarendon Press.
- Cookson, J. (2010). A critique of the specialty certificate examinations of the Federation of Royal Colleges of Physicians of the U.K. *Clinical Medicine (London, England)*, 10(2), 141–4. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/20437984>
- Cote, J. A., & Buckley, R. M. (1987). Estimating Trait , Method and Error Variance: Generalizing Across 70 Construct Validation Studies. *Journal of Marketing*, 24(3), 315–318.
- Cottingham, A. H., Suchman, A. L., Litzelman, D. K., Frankel, R. M., Mossbarger, D. L., Williamson, P. R., ... Inui, T. S. (2008). Enhancing the informal curriculum of a medical school: A case study in organizational culture change. *Journal of General Internal Medicine*, 23, 715–722. doi:10.1007/s11606-008-0543-y
- Coumarbatch, J., Robinson, L., Thomas, R., & Bridge, P. D. (2010). Strategies for identifying students at risk for USMLE step 1 failure. *Family Medicine*, 42(2), 105–10. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/20135567>

- Cronbach, L. J. (1970). *Essentials of Psychological Testing* (3rd ed.). Harper International Edition, Harper & Row.
- Cronbach, L. J. (1990). *Essentials of Psychological Testing* (5th ed.). New York: Harper Collins Publishers.
- Crowne, D. P., & Marlowe, D. (1964). *The approval motive: studies in evaluative dependence*. New York.
- Cruess, R. L., Cruess, S. R., & Johnston, S. E. (2000). Professionalism: an ideal to be sustained. *Lancet*, 356(9224), 156–9. doi:10.1016/S0140-6736(00)02458-2
- Dacre, J., Besser, M., & White, P. (2003). MRCP(UK) PART 2 Clinical Examination (PACES): a review of the first four examination sessions (June 2001 - July 2002). *Clinical Medicine (London, England)*, 3(5), 452–9. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/14601946>
- Dacre, J. E., & Mucklow, J. (2010). A critique of the specialty certificate examinations of the Federation of Royal Colleges of Physicians of the UK. *Journal of the Royal College of Physicians*, 519–520. Retrieved from <http://www.ingentaconnect.com/content/rcop/cm/2010/00000010/00000002/art00013>
- Dauphinee, W. D. (2005). Self regulation must be made to work. *BMJ (Clinical Research Ed.)*, 330(7504), 1385–7. doi:10.1136/bmj.330.7504.1385
- Davies, H., Archer, J., Southgate, L., & Norcini, J. J. (2009). Initial evaluation of the first year of the Foundation Assessment Programme. *Medical Education*, 43(1), 74–81. doi:10.1111/j.1365-2923.2008.03249.x
- DeMaio, T. J. (1984). Social desirability and survey measurement: A review. *Surveying Subjective Phenomena*, 2, 257–281.
- Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. ...*, 39(1), 1–38. Retrieved from <http://www.jstor.org/stable/10.2307/2984875>
- Dewhurst, N. G., McManus, I. C., Mollon, J., Dacre, J. E., & Vale, A. J. (2007). Performance in the MRCP(UK) Examination 2003-4: analysis of pass rates of UK graduates in relation to self-declared ethnicity and gender. *BMC Medicine*, 5, 8. doi:10.1186/1741-7015-5-8
- Doherty, E. M., & Nugent, E. (2011). Personality factors and medical training: a review of the literature. *Medical Education*, 45(2), 132–40. doi:10.1111/j.1365-2923.2010.03760.x
- Donnon, T., Paolucci, E. O., & Violato, C. (2007). The predictive validity of the MCAT for medical school performance and medical board licensing examinations: a meta-analysis of the published research. *Academic Medicine : Journal of the Association of American Medical Colleges*, 82(1), 100–6. doi:10.1097/01.ACM.0000249878.25186.b7

- Downing, S. M. (2003). Validity: on meaningful interpretation of assessment data. *Medical Education*, 37(9), 830–7. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/14506816>
- Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Medical Education*, 38(9), 1006–1012. doi:10.1046/j.1365-2929.2004.01932.x
- Dreyfus, H., & Dreyfus, S. E. (1986). *Mind over machine: the power of human intuition and expertise in the era of the computer*. Oxford: Basil Blackwell.
- Dunleavy, D. M., Kroopnick, M. H., Dowd, K. W., Searcy, C. A., & Zhao, X. (2013). The predictive validity of the MCAT exam in relation to academic performance through medical school: a national cohort study of 2001-2004 matriculants. *Academic Medicine : Journal of the Association of American Medical Colleges*, 88(5), 666–71. doi:10.1097/ACM.0b013e3182864299
- Durbin, J., Watson, G. S., & Durbin, B. Y. J. (1950). Testing for Serial Correlation in Least Squares Regression: I. *Biometrika*, 37(3/4), 409–428.
- Elder, A. T., Mcalpine, L., Bateman, N., Dacre, J., Kopelman, P., & Mcmanus, I. C. (2011). Changing PACES : developments to the examination in 2009. *Clinical Medicine*, 11(3), 2009–2012.
- Emery, J. L., & Bell, J. F. (2009). The predictive validity of the BioMedical Admissions Test for pre-clinical examination performance. *Medical Education*, 43(6), 557–64. doi:10.1111/j.1365-2923.2009.03367.x
- Enders, C., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods*, 12(2), 121–138. doi:10.1037/1082-989X.12.2.121
- Enns, M. W., Cox, B. J., Sareen, J., & Freeman, P. (2001). Adaptive and maladaptive perfectionism in medical students : a longitudinal investigation, 1034–1042.
- Epstein, R. M. (2002). Defining and assessing professional competence. *JAMA: The Journal of the American Medical Association*, 287(2). Retrieved from <http://jama.ama-assn.org/content/287/2/226.short>
- Eraut, M. (2003). *Developing Professional Knowledge and Competence*. Falmer Press, Taylor&Francis Group.
- Esmail, A., & Roberts, C. (2013). Academic performance of ethnic minority candidates and discrimination in the MRCGP examinations between 2010 and 2012: analysis of data. *BMJ: British Medical Journal*, 366(September), 1–10. doi:10.1136/bmj.f5662
- Eva, K. W., Reiter, H. I., Rosenfeld, J., Trinh, K., Wood, T. J., & Norman, G. R. (2012). Association between a medical school admission process using the multiple mini-interview and national licensing examination scores. *JAMA : The Journal of the American Medical Association*, 308(21), 2233–40. doi:10.1001/jama.2012.36914



- Evetts, J. (2003). The Sociological Analysis of Professionalism: Occupational Change in the Modern World. *International Sociology*, 18(2), 395–415. doi:10.1177/0268580903018002005
- Evetts, J. (2006). Introduction: Trust and Professionalism: Challenges and Occupational Changes. *Current Sociology*, 54(4), 515–531. doi:10.1177/0011392106065083
- Feltovich, P. J., Prietula, M. J., & Ericsson, K. A. (2006). Studies of Expertise from Psychological Perspectives. In P. J. Feltovich, K. A. Ericsson, N. Charness, & R. R. Hoffman (Eds.), *The Cambridge Book of Expertise and Expert Performance* (pp. 41–68). Cambridge University Press.
- Ferguson, E., James, D., & Madeley, L. (2002). Factors associated with success in medical school: systematic review of the literature. *British Medical*, 324(April), 952–957.
- Fernandez, N., Dory, V., Ste-Marie, L.-G., Chaput, M., Charlin, B., & Boucher, A. (2012). Varying conceptions of competence: an analysis of how health sciences educators define competence. *Medical Education*, 46(4), 357–65. doi:10.1111/j.1365-2923.2011.04183.x
- Field, A. (2009). *Discover statistics using SPSS* (3rd Editio.). Sage Publications.
- Finn, G., Sawdon, M., Clipsham, L., & McLachlan, J. C. (2009). Peer estimation of lack of professionalism correlates with low Conscientiousness Index scores. *Medical Education*, 43(10), 960–7. doi:10.1111/j.1365-2923.2009.03453.x
- Fleming, P. R., Manderson, W. G., Matthews, M. B., Sanderson, P. H., & Stokes, J. F. (1974). Evaluation of an Examination: M.R.C.P. (U.K.). *Medical Education*, 2(5910), 99–107.
- Fredricksen, N., Mislevy, R. J., & Bejar, I. I. (1993). *Test Theory for a New Generations of Tests*. Hillsdale, New Jersey: Lawrence Erlbaum Associates Inc.
- Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology. General*, 141(1), 2–18. doi:10.1037/a0024338
- Gaito, J. (1959). Non-parametric methods in psychological research. *Psychological Reports*, 5, 115–125. Retrieved from <http://www.amsciepub.com/doi/pdf/10.2466/pr0.1959.5.g.115>
- Galbraith, R. (1990). The radial plot: graphical assessment of spread in ages. *International Journal of Radiation Applications and ...*, 17(3), 207–214. Retrieved from <http://www.sciencedirect.com/science/article/pii/135901899090036W>
- Galbraith, R. F. (1988). A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine*, 7(8), 889–94. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/3413368>
- Gandy, R., Herial, N., Khuder, S., & Metting, P. (2008). Use of Curricular and Extracurricular Assessments to Predict Performance on the United States Medical Licensing

- Examination (USMLE) Step 1: A Multi-Year Study. *Learning Assistance Review*, 13(2). Retrieved from <http://eric.ed.gov/?id=EJ818229>
- Gill, D., & Griffin, A. (2010). Good Medical Practice: What are we trying to say? Textual analysis using tag clouds. *Medical Education*, 44(0), 316–322. doi:10.1111/j.1365-2923.2009.03588.x
- Glaros, A. G., Hanson, A., & Adkison, L. R. (2014). Early Prediction of Medical Student Performance on Initial Licensing Examinations. *Medical Science Educator*, (May).
- GMC. (1867). REGULATIONS OF THE GENERAL MEDICAL COUNCIL AND MEDICAL LICENSING BODIES (Session 1867-68). *British Medical Journal*, 218–234.
- GMC. (1879). REGULATIONS OF THE GENERAL MEDICAL COUNCIL AND MEDICAL LICENSING BODIES (Session 1879-80). *British Medical Journal*, 397–438.
- GMC. (1880). REGULATIONS OF THE GENERAL MEDICAL COUNCIL AND MEDICAL LICENSING BODIES (Session 1880-81). *British Journal of Anaesthesia*, (September), 415–454.
- GMC. (1963). *Functions, Procedure, and Disciplinary Jurisdiction* (1st edition). London: GMC. Retrieved from [http://www.gmc-uk.org/func\\_proced\\_and\\_dis\\_1963.pdf\\_25416891.pdf](http://www.gmc-uk.org/func_proced_and_dis_1963.pdf_25416891.pdf)
- GMC. (1995). *Good Medical Practice* (1st edition). London: GMC. Retrieved from [http://www.gmc-uk.org/good\\_medical\\_practice\\_oct\\_1995.pdf\\_25416576.pdf](http://www.gmc-uk.org/good_medical_practice_oct_1995.pdf_25416576.pdf)
- GMC. (2009). Licensing. Information for doctors' employers and other organisations. General Medical Council. Retrieved from [http://www.gmc-uk.org/Resource\\_pack\\_\\_\\_A4\\_brochure\\_\\_\\_v18\\_\\_\\_FINAL\\_.pdf\\_34007400.pdf](http://www.gmc-uk.org/Resource_pack___A4_brochure___v18___FINAL_.pdf_34007400.pdf)
- GMC. (2011). List of Registered Medical Practitioners. *www.gmc-uk.org*. Retrieved from <http://www.gmc-uk.org/doctors/register/LRMP.asp>
- GMC. (2012a). *2011 Annual Statistics. Fitness to Practice*. Retrieved from [http://www.gmc-uk.org/GMC\\_Annual\\_Statistics\\_2011\\_2.pdf\\_50323977.pdf](http://www.gmc-uk.org/GMC_Annual_Statistics_2011_2.pdf_50323977.pdf)
- GMC. (2012b). 2012 The Good Medical Practice Framework for appraisal and revalidation.
- GMC. (2012c). Publication and disclosure policy. General Medical Council. Retrieved from [http://www.gmc-uk.org/Publication\\_and\\_disclosure\\_policy.pdf\\_36609763.pdf](http://www.gmc-uk.org/Publication_and_disclosure_policy.pdf_36609763.pdf)
- GMC. (2013a). Good medical practice.
- GMC. (2013b). The Medical Register. Retrieved from [http://www.gmc-uk.org/doctors/medical\\_register.asp](http://www.gmc-uk.org/doctors/medical_register.asp)
- GMC. (2013c). The role of the GMC. Retrieved from <http://www.gmc-uk.org/about/role.asp>
- GMC. (2014a). GMC LRMP statistics. *www.gmc-uk.org*. Retrieved September 4, 2014, from [http://www.gmc-uk.org/doctors/register/search\\_stats.asp](http://www.gmc-uk.org/doctors/register/search_stats.asp)

- GMC. (2014b). Guidance on GMC thresholds for the Fitness to Practice investigations.
- GMC. (2014c). Registration , licence to practise and revalidation legislation. *www.gmc-uk.org*. Retrieved June 13, 2014, from [http://www.gmc-uk.org/about/legislation/registration\\_legislation.asp](http://www.gmc-uk.org/about/legislation/registration_legislation.asp)
- Goldstein, E. A., Maestas, R. R., Fryer-Edwards, K., Wenrich, M. D., Oelschlager, A.-M. A., Baernstein, A., & Kimball, H. R. (2006). Professionalism in medical education: an institutional challenge. *Academic Medicine : Journal of the Association of American Medical Colleges*, 81(10), 871–876. doi:10.1097/01.ACM.0000238199.37217.68
- Gonnella, J. S., Erdmann, J. B. J. B., & Hojat, M. (2004). An empirical study of the predictive validity of number grades in medical school using 3 decades of longitudinal data: implications for a grading system. *Medical Education*, 38(4), 425–434. doi:10.1046/j.1365-2923.2004.01774.x
- Grosch, E. N. (2006). Does specialty board certification influence clinical outcomes? *Journal of Evaluation in Clinical Practice*, 12(5), 473–81. doi:10.1111/j.1365-2753.2006.00556.x
- Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional Psychology*, 11(3), 385–398. Retrieved from <http://psycnet.apa.org/journals/pro/11/3/385/>
- Gulliksen, H. (1950). *Theory of Mental Tests* (1st editio.). New York: John Wiley & Sons.
- Haas, J. S., Orav, E. J., & Goldman, L. (1995). The relationship between physicians' qualifications and experience and the adequacy of prenatal care and low birthweight. *American Journal of Public Health*, 85(8 Pt 1), 1087–91. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1615802&tool=pmcentrez&rendertype=abstract>
- Hafferty, F., & Franks, R. (1994). The hidden curriculum, ethics teaching, and the structure of medical education. *Academic Medicine*, 69(11), 861–871. Retrieved from [http://journals.lww.com/academicmedicine/Abstract/1994/11000/The\\_hidden\\_curriculum\\_ethics\\_teaching\\_and\\_the.1.aspx](http://journals.lww.com/academicmedicine/Abstract/1994/11000/The_hidden_curriculum_ethics_teaching_and_the.1.aspx)
- Haight, S. J., Chibnall, J. T., Schindler, D. L., & Slavin, S. J. (2012). Associations of Medical Student Personality and Health / Wellness Characteristics With Their Medical School Performance Across the Curriculum. *Academic Medicine*, 87(4), 476–485. doi:10.1097/ACM.0b013e318248e9d0
- Hammer, D. (2000). Professional attitudes and behaviors: the “A’s and B’s” of professionalism. *American Journal of Pharmaceutical Education*, 64, 455–464. Retrieved from <http://www.aacp.org/resources/studentaffairspersonnel/studentaffairspolicies/documents/asandbsofprofessionalism.pdf>
- Hammond, S. M., O’Rourke, M., Kelly, M., Bennett, D., & O’Flynn, S. (2012). A psychometric appraisal of the DREEM. *BMC Medical Education*, 12(1), 2. doi:10.1186/1472-6920-12-2

- Haq, I., Higham, J., Morris, R., & Dacre, J. (2005). Effect of ethnicity and gender on performance in undergraduate medical examinations. *Medical Education*, 39(11), 1126–8. doi:10.1111/j.1365-2929.2005.02319.x
- Hargreaves, D. H. (1994). The new professionalism: The synthesis of professional and institutional development. *Teaching and Teacher Education*, 10(4), 423–438. doi:10.1016/0742-051X(94)90023-X
- Hausknecht, J. P., Trevor, C. O., & Farr, J. L. (2002). Retaking ability tests in a selection setting: implications for practice effects, training performance, and turnover. *The Journal of Applied Psychology*, 87(2), 243–254. doi:10.1037/0021-9010.87.2.243
- Hawkins, R. E., Margolis, M. J., Durning, S. J., & Norcini, J. J. (2010). Constructing a validity argument for the mini-Clinical Evaluation Exercise: a review of the research. *Academic Medicine : Journal of the Association of American Medical Colleges*, 85(9), 1453–61. doi:10.1097/ACM.0b013e3181eac3e6
- Hawtin, K. E., Williams, H. R. T., McKnight, L., & Booth, T. C. (2014). Performance in the FRCR (UK) Part 2B examination: Analysis of factors associated with success. *Clinical Radiology*, 69(7), 750–757. doi:10.1016/j.crad.2014.03.007
- Hedges, L. V. (1989). An unbiased correction for sampling error in validity generalization studies. *Journal of Applied Psychology*, 74(3), 469–477. doi:10.1037//0021-9010.74.3.469
- Hemphill, J. F. (2003). Interpreting the Magnitudes of Correlation Coefficients. *American Psychologist*, 58(1), 78–79. doi:10.1037/0003-066X.58.1.78
- Herling, R. W. (2000). Operational Definitions of Expertise and Competence. *Advances in Developing Human Resources*, 2(1), 8–21. doi:10.1177/152342230000200103
- Hickson, D. J., & Thomas, M. W. (1969). Professionalization in Britain: A preliminary measurement. *Sociology*, 3(1), 37–53. doi:0803973233
- Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11), 1539–58. doi:10.1002/sim.1186
- Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ (Clinical Research Ed.)*, 327(7414), 557–60. doi:10.1136/bmj.327.7414.557
- Hill, A. B. (1965). The Environment and Disease: Association or Causation. *Proceedings of the Royal Society of Medicine*, 58, 295–300. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1898525&tool=pmcentrez&rendertype=abstract>
- Hilton, S. R., & Slotnick, H. B. (2005). Proto-professionalism: how professionalisation occurs across the continuum of medical education. *Medical Education*, 39(1), 58–65. doi:10.1111/j.1365-2929.2004.02033.x

- Hodges, B. D., Ginsburg, S., Cruess, R., Cruess, S., Delport, R., Hafferty, F., ... Wade, W. (2011). Assessment of professionalism: recommendations from the Ottawa 2010 Conference. *Medical Teacher*, 33(5), 354–63. doi:10.3109/0142159X.2011.577300
- Hofstee, W. K. B., Berge, J., & Hendriks, A. (1998). How to score questionnaires. *Personality and Individual Differences*, 25(5), 897–909. Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S0191886998000865>
- Hojat, M., Paskin, D. L. D. L., Callahan, C. a, Nasca, T. J. T. J., Louis, D. Z. D. Z., Veloski, J., ... Gonnella, J. S. (2007). Components of postgraduate competence: analyses of thirty years of longitudinal data. *Medical Education*, 41(10), 982–989. doi:10.1111/j.1365-2923.2007.02841.x
- Holmboe, E. S., Lipner, R., & Greiner, A. (2008). Assessing quality of care: knowledge matters. *JAMA : The Journal of the American Medical Association*, 299(3), 338–40. doi:10.1001/jama.299.3.338
- Holmboe, E. S., Sherbino, J., Long, D. M., Swing, S. R., & Frank, J. R. (2010). The role of assessment in competency-based medical education. *Medical Teacher*, 32(8), 676–82. doi:10.3109/0142159X.2010.500704
- Hood, S. B. (2009). Validity in Psychological Testing and Scientific Realism. *Theory & Psychology*, 19(4), 451–473. doi:10.1177/0959354309336320
- Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: guidelines for determining model fit. *Articles*, 6(1), 53–60. Retrieved from <http://arrow.dit.ie/cgi/viewcontent.cgi?article=1001&context=buschmanart>
- HSCIC. (2013a). Hospital Episode Statistics, Admitted Patient Care, England 2012-13. [www.hscic.gov.uk](http://www.hscic.gov.uk). Retrieved from <http://www.hscic.gov.uk/searchcatalogue?productid=13264&q=title:“Hospital+Episode+Statistics,+Admitted+patient+care+-+England”&sort=Relevance&size=10&page=1#top>
- HSCIC. (2013b). NHS written complaints 2011-2012.pdf. [www.hscic.gov.uk](http://www.hscic.gov.uk). Retrieved from <http://www.hscic.gov.uk/catalogue/PUB07197>
- Humphrey, C., Hickman, S., & Gulliford, M. C. (2011). Place of medical qualification and outcomes of UK General Medical Council “fitness to practise” process: cohort study. *BMJ: British Medical ...*, 1–9. doi:10.1136/bmj.d1817
- Hunt, M. T., & Glucksman, M. E. (1991). A review of 7 years of complaints in an inner-city accident and emergency department. *Archives of Emergency Medicine*, 8(1), 17–23. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1285728&tool=pmcentrez&rendertype=abstract>
- Hunter, J. E., & Schmidt, F. L. (2004). Meta-Analysis of Correlations Corrected Individually for Artifacts. In *Methods of Meta-Analysis. Correcting errors and bias in research findings*. (Second Edi.). Newbury Park, California: Sage Publications Inc.

- Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *The Journal of Applied Psychology*, 91(3), 594–612. doi:10.1037/0021-9010.91.3.594
- Hunter, M., & May, R. (1993). Some myths concerning parametric and nonparametric tests. *Canadian Psychology/Psychologie ...*, 55(2), 384–389. Retrieved from <http://psycnet.apa.org/journals/cap/34/4/384/>
- Hutchinson, L., Aitken, P., & Hayes, T. (2002). Are medical postgraduate certification processes valid? A systematic review of the published evidence. *Medical Education*, 36(1), 73–91. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2923.2002.01120.x/full>
- IBM Corp. (2010). SPSS Statistics for Windows. Armonk, NY: IBM Corp.
- Ipsos Mori. (2015). Trust in professions. [www.ipsos-mori.com](http://www.ipsos-mori.com). Retrieved March 28, 2015, from <https://www.ipsos-mori.com/researchpublications/researcharchive/15/Trust-in-Professions.aspx>
- Irvine, D. (1999). The performance of doctors: the new professionalism. *The Lancet*, 353, 1–4. Retrieved from <http://www.thelancet.com/pdfs/journals/lancet/PIIS0140673699911601.pdf>
- Irvine, D. (2001). Doctors in the UK: Their new professionalism and its regulatory framework. *Lancet*. doi:10.1016/S0140-6736(01)06800-3
- Jacobs, A. K. (2005). Rebuilding an enduring trust in medicine: a global mandate: presidential address American Heart Association Scientific Sessions 2004. *Circulation*, 111(25), 3494–8. doi:10.1161/CIRCULATIONAHA.105.166277
- Jagpal, H. S. (1982). Multicollinearity in Structural Equation Models With Unobservable Variables. *Journal of Marketing Research*, 19(4), 431–439.
- Jha, V., Bekker, H. L., Duffy, S. R., & Roberts, T. E. (2007). A systematic review of studies assessing and facilitating attitudes towards professionalism in medicine. *Medical Education*, 41(8), 822–9. doi:10.1111/j.1365-2923.2007.02804.x
- Jones, L., & Green, J. (2006). Shifting discourses of professionalism: A case study of general practitioners in the United Kingdom. *Sociology of Health and Illness*, 28(7), 927–950. doi:10.1111/j.1467-9566.2006.00513.x
- Julian, E. R. (2005, October). Validity of the Medical College Admission Test for predicting medical school performance. *Academic Medicine: Journal of the Association of American Medical Colleges*. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/16186610>
- Kadzombe, E. A., & Coals, J. (1992). Complaints against doctors in an accident and emergency department: a 10-year analysis. *Archives of Emergency Medicine*, 9(2), 134–42. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1285850&tool=pmcentrez&rendertype=abstract>

- Kaffash, J. (2012). RCGP faces legal threat over international GP trainee failure rates. *Pulse Today*. Retrieved from <http://www.pulsetoday.co.uk/practice-business/practice-topics/education/rcgp-faces-legal-threat-over-international-gp-trainee-failure-rates/20000852.article#.US4Dd4Yko4x>
- Kaffash, J. (2013). Lawyers give RCGP three weeks to sort CSA , or face legal action. *Pulse Today*. Retrieved from <http://www.pulsetoday.co.uk/practice-business/practice-topics/education/lawyers-give-rcgp-three-weeks-to-sort-csa-or-face-legal-action/20001651.article#.US4CxYYko4w>
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319–342. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2001.tb01130.x/abstract>
- Kaplan, R. M., & Saccuzzo, D. P. (1993). *Psychological Testing: principles, applications and issues*. (3rd ed.). Brooks/Cole Publishing Company.
- Kelly, M., O'Flynn, S., McLachlan, J. C., & Sawdon, M. A. (2012). The clinical conscientiousness index: a valid tool for exploring professionalism in the clinical undergraduate setting. *Academic Medicine : Journal of the Association of American Medical Colleges*, 87(9), 1218–24. doi:10.1097/ACM.0b013e3182628499
- Kelly, J. V., & Hellinger, F. J. (1986). Physician and Hospital Factors Associated With Mortality of Surgical Patients. *Medical Care*, 24(9), 785–800.
- Kernohan, R. J. (1962). Editor's Letter Box. *British Medical Journal*, (3), 1962–1962.
- Kobrin, J. L., Kim, Y., & Sackett, P. R. (2011). Modeling the Predictive Validity of SAT Mathematics Items Using Item Characteristics. *Educational and Psychological Measurement*, 72(1), 99–119. doi:10.1177/0013164411404410
- Koczwara, A., Patterson, F., Zibarras, L., Kerrin, M., Irish, B., & Wilkinson, M. (2012). Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection. *Medical Education*, 46(4), 399–408. doi:10.1111/j.1365-2923.2011.04195.x
- Korsch, B., Gozzi, E., & Francis, V. (1968). Gaps in doctor-patient communication I. Doctor-patient interaction and patient satisfaction. *Pediatrics*, 42(5), 855. Retrieved from <http://pediatrics.aappublications.org/content/42/5/855.short>
- Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing Multilevel Modling*. Sage Publications Ltd.
- Landon, B. E., Normand, S.-L. T., Blumenthal, D., & Daley, J. (2003). Physician Clinical Performance Assessment. Prospects and Barriers. *JAMA : The Journal of the American Medical Association*, 290(9), 1183–9. doi:10.1001/jama.290.9.1183
- Lanier, D. C., Roland, M., Burstin, H., & Knottnerus, J. A. (2003). Doctor performance and public accountability. *Lancet*, 362(9393), 1404–8. doi:10.1016/S0140-6736(03)14638-7

- Lau, F. L. (2000). Can communication skills workshops for emergency department doctors improve patient satisfaction? *Journal of Accident & Emergency Medicine*, 17(4), 251–3. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1725405&tool=pmcentrez&rendertype=abstract>
- Lee, J. M., Kleinbaum, D., Diener-west, M., Grimley, D. M., Wendy, L., Sullivan, L., ... Reports, P. H. (2008). TEACHING EXCELLENCE IN PUBLIC HEALTH: A CALL TO ACTION. *Public Health Reports*, 123(3), 405–407.
- Lempp, H., & Seale, C. (2004). The hidden curriculum in undergraduate medical education : qualitative study of medical students ' perceptions of teaching, 329(October). doi:10.1136/bmj.38202.667130.55
- Levy, J. B., Mohanaruban, A., & Smith, D. (2011a). The relationship between performance in work place based assessments, MRCP exam and outcome from core medical training. *Medical Education*, 45(Supplement2), 5. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2923.2011.04093.x/pdf>
- Levy, J. B., Mohanaruban, A., & Smith, D. (2011b). Utility of Workplace Based Assessments in Postgraduate Core Medical Training. *Medical Education*, 45(Supplement s2), 4. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2923.2011.04093.x/pdf>
- Li, H. (2003). The Resolution of Some Paradoxes Related to Reliability and Validity. *Journal of Educational and Behavioral Statistics*, 28(2), 89–95. doi:10.3102/10769986028002089
- Lievens, F., & Coetsier, P. (2002). Situational Tests in Student Selection : An Examination of Predictive Validity , Adverse Impact , and Construct Validity. *International Journal of Selection and Assessment*, 10(4), 245–257.
- Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *The Journal of Applied Psychology*, 96(5), 927–40. doi:10.1037/a0023496
- Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *The Journal of Applied Psychology*, 97(2), 460–8. doi:10.1037/a0025741
- Linder, J. A., Ma, J., Bates, D. W., Middleton, B., & Stafford, R. S. (2007). Electronic health record use and the quality of ambulatory care in the United States. *Archives of Internal Medicine*, 167(13), 1400–5. doi:10.1001/archinte.167.13.1400
- Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36(8), 437. doi:10.3102/0013189X0731



- Little, R. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404), 1198–1203. Retrieved from <http://amstat.tandfonline.com/doi/full/10.1080/01621459.1988.10478722>
- Little, R. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association*. Retrieved from <http://www.tandfonline.com/doi/full/10.1080/01621459.1992.10476282>
- Lynch, D. C., Surdyk, P. M., & Eiser, A. R. (2004). Assessing professionalism: a review of the literature. *Medical Teacher*, 26(4), 366–73. doi:10.1080/01421590410001696434
- Lyons, L. (1998). Meta-analysis: Methods of accumulating results across research domains. *www.lyonsmorris.com*. Retrieved January 8, 2014, from <http://www.lyonsmorris.com/MetaA/macalc/MAPaper.pdf>
- Ma, J., & Stafford, R. (2005). Quality of US outpatient care: temporal changes and racial/ethnic disparities. *Archives of Internal Medicine*, 165, 1354–1361. Retrieved from <http://archinte.ama-assn.org/cgi/reprint/165/12/1354.pdf>
- Markus, K. (1998). Science, measurement, and validity: Is completion of Samuel Messick's synthesis possible? *Social Indicators Research*, 45, 7–34. Retrieved from <http://www.springerlink.com/index/M0J3T92723TH2J5V.pdf>
- Martimianakis, M. A., Maniate, J. M., & Hodges, B. D. (2009). Sociological interpretations of professionalism. *Medical Education*, 43(ext 28390), 829–837. doi:10.1111/j.1365-2923.2009.03408.x
- Matton, N., Vautier, S., & Raufaste, É. (2009). Situational effects may account for gain scores in cognitive ability testing: A longitudinal SEM approach. *Intelligence*, 37, 412–421. doi:10.1016/j.intell.2009.03.011
- Maudsley, G., & Strivens, J. (2000). "Science", "critical thinking" and "competence" for Tomorrow's Doctors. A review of terms and concepts. *Medical Education*, 34(1), 53–60. doi:10.1046/j.1365-2923.2000.00428.x
- Maurer, T. J., Alexander, R. A., Callahan, C. M., Bailey, J. J., & Dambrot, F. H. (1991). Methodological and psychometric issues in setting cutoff scores using the Angoff method. *Personnel Psychology*, 44, 235–262. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1111/j.1744-6570.1991.tb00958.x/abstract>
- McCaskill, Q. E., Kirk, J. J., Barata, D. M., Wludyka, P. S., Zenni, E. a, & Chiu, T. T. (2007). USMLE step 1 scores as a significant predictor of future board passage in pediatrics. *Ambulatory Pediatrics : The Official Journal of the Ambulatory Pediatric Association*, 7(2), 192–5. doi:10.1016/j.ambp.2007.01.002
- McGaghie, W. C., Cohen, E. R., & Wayne, D. B. (2011). Are United States Medical Licensing Exam Step 1 and 2 scores valid measures for postgraduate medical residency selection decisions? *Academic Medicine : Journal of the Association of American Medical Colleges*, 86(1), 48–52. doi:10.1097/ACM.0b013e3181ffacdb

- McHugh, M. (2008). Standard error: meaning and interpretation. *Biochemia Medica*, 18(1), 7–13. Retrieved from [http://hrcak.srce.hr/index.php?show=clanak&id\\_clanak\\_jezik=31652](http://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=31652)
- McKinstry, B., Walker, J., Blaney, D., Heaney, D., & Begg, D. (2004). Do patients and expert doctors agree on the assessment of consultation skills? A comparison of two patients consultation assessment scales with the video component of the MRCGP. *Family Practice*, 21(1), 75–80. doi:10.1093/fampra/cmh116
- McLachlan, J. C. (2010). Measuring conscientiousness and professionalism in undergraduate medical students. *The Clinical Teacher*, 7(1), 37–40. doi:10.1111/j.1743-498X.2009.00338.x
- McLachlan, J. C., Finn, G., & Macnaughton, J. (2009). The conscientiousness index: a novel tool to explore students' professionalism. *Academic Medicine: Journal of the Association of American Medical Colleges*, 84(5), 559–65. doi:10.1097/ACM.0b013e31819fb7ff
- McManus, I. C., Dewberry, C., Nicholson, S., Dowell, J. S., Woolf, K., & Potts, H. W. W. (2013). Construct-level predictive validity of educational attainment and intellectual aptitude tests in medical student selection: meta-regression of six UK longitudinal studies. *BMC Medicine*, 11, 243. doi:10.1186/1741-7015-11-243
- McManus, I. C., Elder, A. T., de Champlain, A., Dacre, J. E., Mollon, J., & Chis, L. (2008). Graduates of different UK medical schools show substantial differences in performance on MRCP(UK) Part 1, Part 2 and PACES examinations. *BMC Medicine*, 6, 5. doi:10.1186/1741-7015-6-5
- McManus, I. C., Ferguson, E., Wakeford, R., Powis, D., & James, D. (2011). Predictive validity of the Biomedical Admissions Test: an evaluation and case study. *Medical Teacher*, 33(1), 53–7. doi:10.3109/0142159X.2010.525267
- McManus, I. C., & Lissauer, T. (2005). Detecting cheating in written medical examinations by statistical analysis of similarity of answers: pilot study. *BMJ*, 330(May), 1064–1066. Retrieved from <http://www.bmj.com/content/330/7499/1064.short>
- McManus, I. C., & Ludka, K. (2012). Resitting a high-stakes postgraduate medical examination on multiple occasions : nonlinear multilevel modelling of performance in the MRCP ( UK ) examinations. *BMC Medicine*, 10, 60. doi:10.1186/1741-7015-10-60
- McManus, I. C., Mollon, J., Duke, O. L., & Vale, J. A. (2005). Changes in standard of candidates taking the MRCP(UK) Part 1 examination, 1985 to 2002: analysis of marker questions. *BMC Medicine*, 3, 13. doi:10.1186/1741-7015-3-13
- McManus, I. C., Mooney-Somers, J., Dacre, J. E., & Vale, J. a. (2003). Reliability of the MRCP(UK) Part I Examination, 1984-2001. *Medical Education*, 37(7), 609–11. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12834418>
- McManus, I. C., & Richards, P. (1996). Final examination performance of medical students from ethnic minorities. *Medical Education*, 30(3), 195–200. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2923.1996.tb00742.x/full>

- McManus, I. C., Richards, P., Winder, B. C., & Sproston, K. A. (1998). Clinical experience , performance in final examinations , and learning style in medical students : prospective study Rapid responses Email alerting service Clinical experience , performance in final examinations , and learning style in medical students :, (June 2006).
- Mcmanus, I. C., Smithers, E., Partridge, P., Keeling, A., & Fleming, P. R. (2003). Learning in practice A levels and intelligence as predictors of medical careers. *British Medical Journal*, 327(June 2005), 139–142.
- McManus, I. C., Thompson, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education*, 6, 42. doi:10.1186/1472-6920-6-42
- McManus, I. C., & Wakeford, R. (2014). PLAB and UK graduates' performance on MRCP (UK) and MRCGP examinations: data linkage study. *BMJ: British Medical Journal*, 348, 1–24. doi:10.1136/bmj.g2621
- McManus, I. C., Woolf, K., & Dacre, J. (2008). The educational background and qualifications of UK medical students from ethnic minorities. *BMC Medical Education*, 8, 21. doi:10.1186/1472-6920-8-21
- McManus, I. C., Woolf, K., Dacre, J., Paice, E., & Dewberry, C. (2013). The Academic Backbone: longitudinal continuities in educational achievement from secondary school and medical school to MRCP(UK) and the specialist register in UK medical students and doctors. *BMC Medicine*, 11(1), 242. doi:10.1186/1741-7015-11-242
- Mechanic, D. (1996). Changing medical organization and the erosion of trust. *The Milbank Quarterly*, 74(2), 171–89. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/8632733>
- Mechanic, D., & Schlesinger, M. (1996). The impact of managed care on patients' trust in medical care and their physicians. *JAMA: The Journal of the American Medical Association*, 275(21), 1693–7. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/8637148>
- Mendoza, J. (1987). Corrections for attenuation and range restriction on the predictor. *Journal of Educational and Behavioral*, 12(3), 282–293. Retrieved from <http://jeb.sagepub.com/content/12/3/282.short>
- Mercer, A., & Puddey, I. B. (2011). Admission selection criteria as predictors of outcomes in an undergraduate medical course: a prospective study. *Medical Teacher*, 33(12), 997–1004. doi:10.3109/0142159X.2011.577123
- Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35(11), 1012–1027. Retrieved from <http://psycnet.apa.org/psycinfo/1981-27017-001>
- Metcalf, N. H. (2012). Testing the test: an analysis of the MRCGP Applied Knowledge Test as an assessment tool. *Education for Primary Care: An Official Publication of the Association of Course Organisers, National Association of GP Tutors, World*

- Organisation of Family Doctors*, 23(1), 13–8. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/22306140>
- Miller, G. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, 65(S63–7.).
- Millerson, G. (1964). *The Qualifying Associations: A Study of Professionalization*. London: Routledge & Kegan Paul.
- Mislevy, R. J. (2004). Can There Be Reliability without “Reliability?” *Journal of Educational and Behavioral Statistics*, 29(2), 241–244. doi:10.3102/10769986029002241
- Mitchell, C., Bhat, S., Herbert, A., & Baker, P. (2011). Workplace-based assessments of junior doctors: do scores predict training difficulties? *Medical Education*, 45(12), 1190–8. doi:10.1111/j.1365-2923.2011.04056.x
- Moses, T., Deng, W., & Zhang, Y. (2010). *The Use of Two Anchors in Nonequivalent Groups With Anchor Test (NEAT) Equating*. Princeton. Retrieved from <https://www.ets.org/Media/Research/pdf/RR-10-23.pdf>
- Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, 23(2), 5–12. Retrieved from <http://edr.sagepub.com/content/23/2/5.short>
- MPTS. (2012). GMC launches new tribunal service for UK doctors. <http://www.mpts-uk.org>. Retrieved April 14, 2015, from <http://www.mpts-uk.org/about/news/1760.asp>
- MPTS. (2014). The role of the MPTS. [www.mpts-uk.org](http://www.mpts-uk.org). Retrieved August 28, 2014, from <http://www.mpts-uk.org/about/1595.asp>
- Munro, N., Denney, M. L., Rughani, A., Foulkes, J., Wilson, A., & Tate, P. (2005). Ensuring reliability in UK written tests of general practice: the MRCGP examination 1998-2003. *Medical Teacher*, 27(1), 37–45. doi:10.1080/01421590400013461
- Nathans, L., Oswald, F., & Nimon, K. (2012). Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research ...*, 17(9), 1–19. Retrieved from <http://www.dspace.rice.edu/handle/1911/71096>
- NHS. (2010a). A Reference Guide for Postgraduate Specialty Training in the UK. NHS.
- NHS. (2010b). NHS e-Portfolio. Retrieved from <https://www.nhseportfolios.org/>
- Nibert, A. T., Young, A., & Adamson, C. (2002). Predicting NCLEX with the HESI Exit Exam. Fourth Annual Validity Study. *CIN: Computers, Informatics, Nursing*, 20(6), 261–267.
- Norcini, J. J. (2003a). Setting standards on educational tests. *Medical Education*, 37(5), 464–9. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12709190>
- Norcini, J. J. (2003b). Work based assessment. *British Medical Journal*, 326(4), 753–755.
- Norcini, J. J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., ... Roberts, T. (2011). Criteria for good assessment: consensus statement and recommendations

- from the Ottawa 2010 Conference. *Medical Teacher*, 33(3), 206–14. doi:10.3109/0142159X.2011.551559
- Norcini, J. J., Blank, L. L., Duffy, F. D., & Fortna, G. S. (2003). The mini-CEX: a method for assessing clinical skills. *Annals of Internal Medicine*, 138(6), 476–481. Retrieved from <http://www.annals.org/content/138/6/476.short>
- Norcini, J. J., Boulet, J. R., Opalek, A., & Dauphinee, W. D. (2014). The relationship between licensing examination performance and the outcomes of care by international medical school graduates. *Academic Medicine : Journal of the Association of American Medical Colleges*, 89(8), 1157–62. doi:10.1097/ACM.0000000000000310
- Norcini, J. J., Kimball, H. R., & Lipner, R. S. (2000). Certification and specialization: do they matter in the outcome of acute myocardial infarction? *Academic Medicine : Journal of the Association of American Medical Colleges*, 75(12), 1193–8. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11112721>
- Norcini, J. J., Lipner, R. S., & Kimball, H. R. (2002). Certifying examination performance and patient outcomes following acute myocardial infarction. *Medical Education*, 36(9), 853–9. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12354248>
- Norcini, J. J., & McKinley, D. (2007). Assessment methods in medical education. *Teaching and Teacher Education*, 23(3), 239–250. doi:10.1016/j.tate.2006.12.021
- Norcini, J. J., & Talati, J. (2009). Assessment, surgeon, and society. *International Journal of Surgery (London, England)*, 7(4), 313–7. doi:10.1016/j.ijsu.2009.06.011
- Norman, G., Eva, K., Brooks, L., & Hamstra, S. (2006). Expertise in Medicine and Surgery. In *The Cambridge Book of Expertise and Expert Performance* (pp. 339–354). New York: Cambridge University Press.
- NPSA. (2011). NRLS Quarterly Data Workbook up to June 2011. Retrieved from <http://www.nrls.npsa.nhs.uk/resources/collections/quarterly-data-summaries/?entryid45=132910>
- Owen, C. (1991). Formal complaints against general practitioners: a study of 1000 cases. *The British Journal of General Practice*, 41, 113–115. Retrieved from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1371624/>
- Pant, M., Nesargikar, P. N., & Cocker, D. M. (2009). Team Assessment Behaviour (TAB) as an assessment tool: A critical evaluation. *BJMP*, 2(3), 35–37. Retrieved from <http://www.bjmp.org/content/team-assessment-behaviour-tab-assessment-tool-critical-evaluation>
- Papadakis, M. A., Arnold, G. K. G., Blank, L. L., Holmboe, E. S., & Lipner, R. S. (2008). Performance during internal medicine residency training and subsequent disciplinary action by state licensing boards. *Annals of Internal Medicine*, 148(11), 869–876. Retrieved from <http://www.annals.org/content/148/11/869.short>

- Papadakis, M. A., Hodgson, C. S., Teherani, A., & Kohatsu, N. D. (2004). Unprofessional Behavior in Medical School Is Associated with Subsequent Disciplinary Action by a State Medical Board. *Academic Medicine*, 79(3), 244–249.
- Papadakis, M. A., Teherani, A., Banach, M. A., Knettler, T. R., Rattner, S. L., Stern, D. T., ... Hodgson, C. S. (2005). Disciplinary Action by Medical Boards and Prior Behavior in Medical School. *The New England Journal of Medicine*, 353, 2673–2682.
- Patterson, H. D., & Thompson, R. (1971). Recovery of Inter-Block Information when Block Sizes are Unequal. *Biometrika*, 58(3), 545. doi:10.2307/2334389
- Pellegrino, E. D. (2002). Professionalism, profession and the virtues of the good physician. *The Mount Sinai Journal of Medicine, New York*, 69(6), 378–84. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12429956>
- Pfadenhauer, M. (2006). Crisis or Decline?: Problems of Legitimation and Loss of Trust in Modern Professionalism. *Current Sociology*, 54(4), 565–578. doi:10.1177/0011392106065088
- Pham, H. H., Schrag, D., Hargraves, J. L., & Bach, P. B. (2005). Delivery of preventive services to older adults by primary care physicians. *JAMA : The Journal of the American Medical Association*, 294(4), 473–81. doi:10.1001/jama.294.4.473
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *The Journal of Applied Psychology*, 88(5), 879–903. doi:10.1037/0021-9010.88.5.879
- Poole, P., Shulruf, B., Rudland, J., & Wilkinson, T. (2012). Comparison of UMAT scores and GPA in prediction of performance in medical school: a national study. *Medical Education*, 46(2), 163–71. doi:10.1111/j.1365-2923.2011.04078.x
- Prideaux, D., Roberts, C., Eva, K., Centeno, A., McCrorie, P., McManus, I. C., ... Wilkinson, D. (2011). Assessment for selection for the health care professions and specialty training: consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, 33(3), 215–23. doi:10.3109/0142159X.2011.551560
- Pugh, D., Hamstra, S. J., Wood, T. J., Humphrey-Murto, S., Touchie, C., Yudkowsky, R., & Bordage, G. (2014). A procedural skills OSCE: assessing technical and non-technical skills of internal medicine residents. *Advances in Health Sciences Education : Theory and Practice*. doi:10.1007/s10459-014-9512-x
- Puhan, G., Sinharay, S., Haberman, S., & Larkin, K. (2010). The Utility of Augmented Subscores in a Licensure Exam: An Evaluation of Methods Using Empirical Data. *Applied Measurement in Education*, 23(3), 266–285. doi:10.1080/08957347.2010.486287
- R Development Core Team. (2005). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.r-project.org>

- Ramos, K., Schafer, S., & Tracz, S. (2003). Validation of the Fresno test of competence in evidence based medicine. *BMJ: British Medical Journal*, 326, 319–321. Retrieved from <http://www.bmj.com/content/326/7384/319?variant=full-text>
- Ramsey, P. G., Carline, J. D., Inui, T. S., Larson, E. B., LoGerfo, J. P., & Wenrich, M. D. (1989). Predictive validity of certification by the American Board of Internal Medicine. *Annals of Internal Medicine*, 110(9), 719–726.
- Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2005). MLwiN. Centre for Multilevel Modelling, University of Bristol.
- RCGP. (2010). RCGP Clinical Skills Assessment. The Royal College of General Practitioners. Retrieved from [http://www.rcgp.org.uk/gp-training-and-exams/mrcgp-exam-overview/~media/Files/GP-training-and-exams/Exams-CSA-powerpoint-num-2-1-\(2\).ashx](http://www.rcgp.org.uk/gp-training-and-exams/mrcgp-exam-overview/~media/Files/GP-training-and-exams/Exams-CSA-powerpoint-num-2-1-(2).ashx)
- RCGP. (2012). Eligibility for MRCGP examinations , number of attempts permitted , and consideration of mitigating circumstances. Retrieved from <http://www.rcgp.org.uk/gp-training-and-exams/~media/Files/GP-training-and-exams/Eligibility for MRCGP examinations number of attempts permitted and consideration of mitigating circumstances.ashx>
- RCGP. (2013a). MRCGP Applied Knowledge Test ( AKT ). Retrieved from <http://www.rcgp.org.uk/gp-training-and-exams/mrcgp-exam-overview/mrcgp-applied-knowledge-test-akt.aspx>
- RCGP. (2013b). MRCGP exam : overview. Retrieved from <http://www.rcgp.org.uk/gp-training-and-exams/mrcgp-exam-overview.aspx>
- RCP. (2005). Doctors in society: medical professionalism in a changing world. *Clinical Medicine (London, England)*, 6(1), 109–13. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/16521367>
- RCP. (2013a). MRCP ( UK ) - Part 2 Written Home. Retrieved from [http://www.mrcpuk.org/Part2/Pages/\\_Home.aspx](http://www.mrcpuk.org/Part2/Pages/_Home.aspx)
- RCP. (2013b). MRCP ( UK ) Part 2 Written Sample Questions. The Royal College of Physicians. Retrieved from <http://www.mrcpuk.org/SiteCollectionDocuments/Part-2-sample-questions.pdf>
- RCP. (2013c). MRCP (UK) Part 1 Exam Format. *www.mrcp.org*. Retrieved from <http://www.mrcpuk.org/Part1/Pages/Part1Format.aspx>
- RCP. (2013d). Part I MRCP. Retrieved from [www.MRCP\(UK\)uk.org/part1](http://www.MRCP(UK)uk.org/part1)
- RCP. (2013e). Specialty Certificate Home Page. Retrieved from <http://www.mrcpuk.org/SCE/Pages/Home.aspx>
- RCP. (2013f). Specialty Certificates- Exam Format. Retrieved from <http://www.mrcpuk.org/SCE/Pages/Home.aspx>

- RCP. (2014a). Exam Pass Marks. [www.mrcp-uk.org](http://www.mrcp-uk.org). Retrieved from <http://www.mrcpuk.org/about-us/research/exam-pass-marks>
- RCP. (2014b). Part 2 MRCP(UK) Format. [www.mrcpuk.org](http://www.mrcpuk.org). Retrieved August 25, 2014, from <http://www.mrcpuk.org/mrcpuk-examinations/part-2/format>
- RCP, & Cuthbertson, L. (2008). A concise History of the MRCP Examination: Celebrating 150 years. *History*. The Royal College of Physicians.
- RCR. (2013a). Final Examination for the Fellowship in Clinical Oncology ( Part A ) Guidance Notes for Candidates. The ROyal College of Radiologists. Retrieved from [http://www.rcr.ac.uk/docs/oncology/pdf/CO2A\\_candidate\\_guidance\\_notes.pdf](http://www.rcr.ac.uk/docs/oncology/pdf/CO2A_candidate_guidance_notes.pdf)
- RCR. (2013b). Final Examination for the Fellowship in Clinical Oncology ( Part B ) Guidance Notes for Candidates. The ROyal College of Radiologists. Retrieved from [http://www.rcr.ac.uk/docs/oncology/pdf/CO2B\\_candidate\\_guidance\\_notes.pdf](http://www.rcr.ac.uk/docs/oncology/pdf/CO2B_candidate_guidance_notes.pdf)
- RCR. (2013c). Regulations for the Examinations for the Fellowship of the Royal College of Radiologists in Clinical Oncology. The ROyal College of Radiologists. Retrieved from [http://www.rcr.ac.uk/docs/oncology/pdf/FellRegsCO\\_20August2012.pdf](http://www.rcr.ac.uk/docs/oncology/pdf/FellRegsCO_20August2012.pdf)
- RCR. (2013d). The Royal College of Radiologists | Final FRCR Examination. Retrieved from <http://www.rcr.ac.uk/content.aspx?PageID=79>
- Reeve, C. L., & Lam, H. (2005). The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score changes. *Intelligence*, 33, 535–549. doi:10.1016/j.intell.2005.05.003
- Reid, R., & Friedberg, M. (2010). Associations between physician characteristics and quality of care. *Archives of Internal Medicine*, 170(16), 1442–1449. Retrieved from <http://archpsyc.jamanetwork.com/article.aspx?articleid=225924>
- Reid, R. R. O., Friedberg, M. M. W., Adams, J. L., McGlynn, E. A., & Mehrotra, A. (2010). Associations between physician characteristics and quality of care. *Archives of Internal Medicine*, 170(16), 1442–9. doi:10.1001/archinternmed.2010.307
- Reinders, M. E., Blankenstein, A. H., van Marwijk, H. W. J., Knol, D. L., Ram, P., van der Horst, H. E., ... van der Vleuten, C. P. M. (2011). Reliability of consultation skills assessments using standardised versus real patients. *Medical Education*, 45(6), 578–84. doi:10.1111/j.1365-2923.2010.03917.x
- Riley, B. (2008). The new MRCGP-what's it all about? *InnovAiT*, 1(1), 49–52. doi:10.1093/innovait/inm013
- Russell, D., Simpson, R., & Rendel, S. (2011). Standardisation of role players for the Clinical Skills Assessment of the MRCGP. *Education for Primary Care : An Official Publication of the Association of Course Organisers, National Association of GP Tutors, World Organisation of Family Doctors*, 22(3), 166–70. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/21640006>



- Rutledge, R., Oller, D. W., Meyer, A., & Johnson, G. J. (1996). A statewide, population-based time-series analysis of the outcome of ruptured abdominal aortic aneurysm. *Annals of Surgery*, 223(5), 492–502; discussion 503–5. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1235169&tool=pmcentrez&rendertype=abstract>
- Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases*, 32, 51–63. doi:10.1016/0021-9681(79)90012-2
- Sackett, P. R., Laczo, R., & Arvey, R. (2002). The effects of range restriction on estimates of criterion interrater reliability: Implications for validation research. *Personnel Psychology*, 55, 807–825. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1111/j.1744-6570.2002.tb00130.x/abstract>
- Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, 85(1), 112–118. doi:10.1037//0021-9010.85.1.112
- SAS Institute Inc. (2004). SAS. Cary, NC, USA: SAS Institute Inc.
- Schlesinger, M. (2002). A loss of faith: the sources of reduced political legitimacy for the American medical profession. *The Milbank Quarterly*, 80(2), 185–235. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12101871>
- Schmidt, F. L., Oh, I.-S., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: model properties and an empirical comparison of differences in results. *The British Journal of Mathematical and Statistical Psychology*, 62(Pt 1), 97–128. doi:10.1348/000711007X255327
- Schumacker, R. E., & Lomax, R. G. (2010). *A Beginner's Guide to Structural Equation Modelling* (Third Edit.). New York: Routledge, part of the Taylor & Francis Group.
- Schwartz, L. R., & Overton, D. T. (1987). Emergency department complaints: a one-year analysis. *Annals of Emergency Medicine*, 16(8), 857–61. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/3619164>
- Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision ...*, 53, 252–266. Retrieved from <http://www.sciencedirect.com/science/article/pii/074959789290064E>
- Sharp, L. K., Bashook, P. G., Lipsky, M. S., Horowitz, S. D., & Miller, S. H. (2002). Specialty board certification and clinical outcomes: the missing link. *Academic Medicine : Journal of the Association of American Medical Colleges*, 77(6), 534–42. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12063199>
- Shaw, K., Cassel, C., Black, C., & Levinson, W. (2009). Shared Medical Regulation in a Time of Increasing Calls for Accountability and Transparency. ... *of the American Medical ...*, 302(18). Retrieved from <http://jama.ama-assn.org/content/302/18/2008.short>
- Siegrist, H. (1994). The professions, state and government in theory and history. In *Governments and Professional Education* (pp. 3–20).

- Silver, N. (2013). *The signal and the noise* (1st ed.). London: Penguin Books.
- Simon, S. R., Volkan, K., Hamann, C., Duffey, C., & Fletcher, S. W. (2002). The relationship between second-year medical students' OSCE scores and USMLE Step 1 scores. *Medical Teacher*, 24(5), 535–9. doi:10.1080/0142159021000012586
- Skaggs, G., & Lissitz, R. W. (1986). IRT Test Equating: Relevant Issues and a Review of Recent Research. *Review of Educational Research*, 56(4), 495–529. doi:10.3102/00346543056004495
- Slowther, A., Lewando Hundt, G., Taylor, R., & Purkis, J. (2009). *Non UK qualified doctors and Good Medical Practice : The experience of working within a different professional framework*. Warwick. Retrieved from [http://www.gmc-uk.org/FINAL\\_GMC\\_Warwick\\_Report.pdf\\_25392230.pdf](http://www.gmc-uk.org/FINAL_GMC_Warwick_Report.pdf_25392230.pdf)
- Smith, A. F., & Greaves, J. D. (2010). Beyond competence: defining and promoting excellence in anaesthesia. *Anaesthesia*, 65(2), 184–91. doi:10.1111/j.1365-2044.2009.06162.x
- Speyer, R., Pilz, W., Van Der Kruis, J., & Brunings, J. W. (2011). Reliability and validity of student peer assessment in medical education: a systematic review. *Medical Teacher*, 33(11), e572–85. doi:10.3109/0142159X.2011.610835
- Statistical Support University of Texas. (2001). Structural Equation Modeling using AMOS : An Introduction Section 1 : Introduction. Austin: University of Texas. Retrieved from <http://www.utexas.edu/its-archive/rc/tutorials/stat/amos/>
- Stauffer, J., & Mendoza, J. (2001). The proper sequence for correcting correlation coefficients for range restriction and unreliability. *Psychometrika*, 66(1), 63–68. Retrieved from <http://www.springerlink.com/index/28G667101018H280.pdf>
- Steenbergen, M. R., & Jones, B. S. (2002). Modeling Multilevel Data Structures. *American Journal of Political Science*, 46(1), 218–237.
- Stern, D. T., Frohna, A. Z., & Gruppen, L. D. (2005). The prediction of professional behaviour. *Medical Education*, 39(1), 75–82. doi:10.1111/j.1365-2929.2004.02035.x
- Sterne, J., Sutton, A., Ioannidis, J., & Terrin, N. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, 1–8. doi:10.1136/bmj.d4002
- Stevens, R. (2001). Public roles for the medical profession in the United States: beyond theories of decline and fall. *The Milbank Quarterly*, 79(3), 327–53, III. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11565160>
- Stoof, A., Martens, R. L., van Merriënboer, J. J. G., & Bastiaens, T. J. (2002). The Boundary Approach of Competence: A Constructivist Aid for Understanding and Using the Concept of Competence. *Human Resource Development Review*, 1(3), 345–365. doi:10.1177/1534484302013005

- Suchman, A. L., Williamson, P. R., Litzelman, D. K., Frankel, R. M., Mossbarger, D. L., & Inui, T. S. (2004). Toward an informal curriculum that teaches professionalism. Transforming the social environment of a medical school. *Journal of General Internal Medicine*, 19, 501–504. doi:10.1111/j.1525-1497.2004.30157.x
- Svensson, L. G. (2006). New Professionalism, Trust and Competence: Some Conceptual Remarks and Empirical Data. *Current Sociology*, 54(4), 579–593. doi:10.1177/0011392106065089
- Swanson, D. B., Case, S. M., Koenig, J., & Killian, C. D. (1996). Preliminary study of the accuracies of the old and new medical college admission tests for predicting performance on USMLE Step 1. *Academic Medicine*, 71(1), S25. Retrieved from [http://journals.lww.com/academicmedicine/Abstract/1996/01000/Preliminary\\_study\\_of\\_the\\_accuracies\\_of\\_the\\_old\\_and.33.aspx](http://journals.lww.com/academicmedicine/Abstract/1996/01000/Preliminary_study_of_the_accuracies_of_the_old_and.33.aspx)
- Swick, H. M. (2000). Toward a normative definition of medical professionalism. *Academic Medicine : Journal of the Association of American Medical Colleges*, 75, 612–616. doi:10.1097/00001888-200006000-00010
- Tamblyn, R., Abrahamowicz, M., Brailovsky, C., Grand'Maison, P., Lescop, J., Norcini, J. J., ... Haggerty, J. (1998). Association between licensing examination scores and resource use and quality of care in primary care practice. *JAMA : The Journal of the American Medical Association*, 280(11), 989–96. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9749481>
- Tamblyn, R., Abrahamowicz, M., Dauphinee, W. D., Hanley, J. A., Norcini, J. J., Girard, N., ... Brailovsky, C. (2002). Association between licensure examination scores and practice in primary care. *JAMA : The Journal of the American Medical Association*, 288(23), 3019–26. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12479767>
- Tamblyn, R., Abrahamowicz, M., Dauphinee, W. D., Wenghofer, E., Jacques, A., Klass, D., ... Girard, N. (2007). Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *JAMA: The Journal of the American Medical Association*, 298(9), 993. Retrieved from <http://jama.ama-assn.org/content/298/9/993.short>
- Tan, L. T., & McAleer, J. J. a. (2008). The introduction of single best answer questions as a test of knowledge in the final examination for the fellowship of the Royal College of Radiologists in Clinical Oncology. *Clinical Oncology (Royal College of Radiologists (Great Britain))*, 20(8), 571–6. doi:10.1016/j.clon.2008.05.010
- Tanizaki, H. (1997). Power comparison of non-parametric tests: Small-sample properties from Monte Carlo experiments. *Journal of Applied Statistics*, 24(5), 603–632. doi:10.1080/02664769723576
- Taylor, D. M., Wolfe, R., & Cameron, P. a. (2002). Complaints from emergency department patients largely result from treatment and communication problems. *Emergency Medicine (Fremantle, W.A.)*, 14(1), 43–9. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11993834>

- Teherani, A., Hodgson, C. S., Banach, M., & Papadakis, M. A. (2005). Domains of unprofessional behavior during medical school associated with future disciplinary action by a state medical board. *Academic Medicine : Journal of the Association of American Medical Colleges*, 80(10 Suppl), S17–S20. doi:10.1097/00001888-200510001-00008
- The Federation of the Royal Colleges of Physicians of the United Kingdom. (2011). *Census of consultant physicians and medical registrars in the UK, 2010*. Retrieved from <http://www.rcplondon.ac.uk/sites/default/files/census-2010.pdf>
- The MathWorks Inc. (2010). Matlab. Natick, Massachusetts: The MathWorks Inc.
- The Royal College of Radiologists. (2011). Clinical Radiology. Retrieved September 26, 2011, from <http://www.rcr.ac.uk/section.aspx?pageID=11>
- The Royal Colleges of Physicians of the United Kingdom. (2011). Regulations and Information for MRCP ( UK ) Candidates. 2011 Edition.
- Tiffin, P. A., McLachlan, J. C., Webster, L., & Nicholson, S. (2014). Comparison of the sensitivity of the UKCAT and A Levels to sociodemographic characteristics: a national study. *BMC Medical Education*, 14, 7. doi:10.1186/1472-6920-14-7
- Tiffin, P., Dowell, J., & McLachlan, J. C. (2012). Widening access to UK medical education for under-represented socioeconomic groups: modelling the impact of the UKCAT in the 2009 cohort. *BMJ: British Medical Journal*, 1805(April), 1–27. doi:10.1136/bmj.e1805
- Tiffin, P., Illing, J., Kasim, A., & McLachlan, J. C. (2014). Annual Review of Competence Progression (ARCP) performance of doctors who passed Professional and Linguistic Assessments Board (PLAB) tests compared with UK medical graduates: national. *BMJ: British Medical Journal*, 348, 1–18. doi:10.1136/bmj.g2622
- Tighe, J., McManus, I. C., Dewhurst, N. G., Chis, L., & Mucklow, J. (2010). The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP(UK) examinations. *BMC Medical Education*, 10, 40. doi:10.1186/1472-6920-10-40
- Tooke, J. S., Ashtiany, S., Carter, D. S., Cole, A., Michael, J., Rashid, A., ... Petty-Saphon, K. (2008). *Aspiring to excellence. Findings and final recommendations of the independent inquiry into Modernising Medical Careers. MMC Inquiry, London, 2008*. Retrieved from [http://www.mmcinquiry.org.uk/Final\\_8\\_Jan\\_08\\_MMC\\_all.pdf](http://www.mmcinquiry.org.uk/Final_8_Jan_08_MMC_all.pdf)
- Tooney, C. Z., & Duval, R. (1993). *Bootstrapping: A Nonparametric Approach to Statistical Inference*. (C. Z. Tooney, Ed.). New York: Sage Publications Inc.
- Tussing, A. D., & Wojtowycz, M. A. (1993). The effect of physician characteristics on clinical behavior: cesarean section in New York State. *Social Science & Medicine*, 37(10), 1251–60. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/8272903>
- UCL Research Ethics Committee. (2011). Exemptions from Ethical Approval. Retrieved from <http://ethics.grad.ucl.ac.uk/exemptions.php>.

- Unwin, E. (n.d.). *Sex differences in the professional performance of doctors practising in the UK*. Phd thesis in preparation. UCL, London.
- Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, 45(2), 576–94. doi:10.3758/s13428-012-0261-6
- Van der Vleuten, C. P. M. (2000). Validity of final examinations in undergraduate medical training. *Educational Research*, 321(11), 1217–1219.
- Van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., & van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher*, 34(3), 205–14. doi:10.3109/0142159X.2012.652239
- Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor package. *Journal Of Statistical Software*, 36(3).
- Viechtbauer, W. (2013). Package “metafor.”
- Waddington, I. (1973). The struggle to reform the Royal College of Physicians, 1767-1771: a sociological analysis. *Medical History*, 17(2), 107–26. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1081439&tool=pmcentrez&rendertype=abstract>
- Wakeford, R. (2011). Who gets struck off? *BMJ*, 7842(December), 1–6. doi:10.1136/bmj.d7842
- Wakeford, R., Denney, M., Ludka-Stempien, K., & Mcmanus, I. C. (2015). Cross-comparison of MRCGP & MRCP(UK) in a database linkage study of 2,284 candidates taking both examinations: assessment of validity and differential performance by ethnicity. *BMC Medical Education*. doi:10.1186/s12909-014-0281-2
- Walsh, M., Bailey, P. H., & Koren, I. (2009). Objective structured clinical evaluation of clinical competence: an integrative review. *Journal of Advanced Nursing*, 65(8), 1584–95. doi:10.1111/j.1365-2648.2009.05054.x
- Wenghofer, E., Klass, D., Abrahamowicz, M., Dauphinee, W. D., Jacques, A., Smees, S., ... Tamblyn, R. (2009). Doctor scores on national qualifying examinations predict quality of care in future practice. *Medical Education*, 43(12), 1166–73. doi:10.1111/j.1365-2923.2009.03534.x
- West, C. P., Huntington, J. L., Huschka, M. M., Novotny, P. J., Sloan, J. a, Kolars, J. C., ... Shanafelt, T. D. (2007). A prospective study of the relationship between medical knowledge and professionalism among internal medicine residents. *Academic Medicine : Journal of the Association of American Medical Colleges*, 82(6), 587–92. doi:10.1097/ACM.0b013e3180555fc5
- Wiberg, M., & Sundström, A. (2009). A comparison of two approaches to correction of restriction of range in correlation analysis, 14(5).

- Wilkinson, D., Zhang, J., & Parker, M. (2011). Predictive validity of the Undergraduate Medicine and Health Sciences Admission Test for medical students' academic performance. *Medical Journal of Australia*, 194(7), 341–344. Retrieved from [https://mjainsight.com.au/system/files/issues/194\\_07\\_040411/wil11056\\_fm.pdf](https://mjainsight.com.au/system/files/issues/194_07_040411/wil11056_fm.pdf)
- Wilkinson, J. R., Crossley, J. G. M., Wragg, A., Mills, P., Cowan, G., & Wade, W. (2008). Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical Education*, 42(4), 364–73. doi:10.1111/j.1365-2923.2008.03010.x
- Willoughby, L., Gammon, L. C., & Jonas, H. S. (1979). Correlates of Clinical Performance During Medical School. *Journal of Medical Education*, 54, 453–460.
- Wofford, M. M., Wofford, J. L., Bothra, J., Kendrick, S. B., Smith, A., & Lichstein, P. R. (2004). Patient complaints about physician behaviors: a qualitative study. *Academic Medicine: Journal of the Association of American Medical Colleges*, 79(2), 134–8. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/14744713>
- Woolf, K., Potts, H., & McManus, I. C. (2011). Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis. *BMJ: British Medical Journal*, 342, 1–14. doi:10.1136/bmj.d901
- Wragg, A., Wade, W., Fuller, G., Cowan, G., & Mills, P. (2003). Assessing the performance of specialist registrars. *Clinical Medicine (London, England)*, 3(2), 131–4. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12737369>
- Yang, Y., & Green, S. B. (2010). A Note on Structural Equation Modeling Estimates of Reliability. *Structural Equation Modeling: A Multidisciplinary Journal*, 17(January 2015), 66–81. doi:10.1080/10705510903438963
- Yates, J., & James, D. (2006). Predicting the “strugglers”: a case-control study of students at Nottingham University Medical School. *BMJ (Clinical Research Ed.)*, 332(7548), 1009–13. doi:10.1136/bmj.38730.678310.63
- Yeung, A., Booth, T. C., Jacob, K., McCoubrie, P., & McKnight, L. (2011). The FRCR 2B examination: a survey of candidate perceptions and experiences. *Clinical Radiology*, 66(5), 412–9. doi:10.1016/j.crad.2010.12.005
- Yeung, A., Booth, T. C., Larkin, T. J., McCoubrie, P., & McKnight, L. (2012). The FRCR 2B oral examination: Is it reliable? *Clinical Radiology*, 1–6. doi:10.1016/j.crad.2012.10.010
- Yoho, R. M., Antonopoulos, K., & Vardaxis, V. (2012). Undergraduate GPAs, MCAT Scores, and Academic Performance the First 2 Years in Podiatric Medical School at Des Moines University. *Journal of the American Podiatric Medical Association*, 102(6), 446–450.
- Zhao, X., Oppler, S., Dunleavy, D., & Kroopnick, M. (2010). Validity of four approaches of using repeaters' MCAT scores in medical school admissions to predict USMLE Step 1 total scores. *Academic Medicine: Journal of the Association of American Medical Colleges*, 85(10 Suppl), S64–7. doi:10.1097/ACM.0b013e3181ed38fc

## Appendix A

### Ethical Approval

To: Ethics <ethics@ucl.ac.uk>  
From: Chris McManus <i.mcmanus@ucl.ac.uk>  
Subject: Re: Is ethics permission required?

Dear Sir John,

Thank you very much for your email, and for the decision, which is much appreciated.

With best wishes

Chris McManus

---

At 12:34 15/02/2011, you wrote:  
Dear Professor McManus

I have considered your research project and concluded that it is indeed exempt from ethics approval. A formal application to the UCL Research Ethics Committee will therefore not be required.

With best wishes

Sir John Birch  
Chair, UCL Research Ethics Committee

---

On 14/02/2011 09:26, Chris McManus wrote:  
To: Sir John Birch, Chair, Ethics Committee, UCL.

Dear Sir John,

I wonder if you could give us a brief opinion as to whether ethics permission is required for some studies we are doing. We believe that they probably are exempt under headings c, d and/or f of the exemptions, but we have been asked to provide confirmation of that.

I am educational advisor to the MRCP(UK) examination, which runs large-scale assessments for doctors in postgraduate training. Professor Jane Dacre of UCL (DoME/ACME) is also Medical Director of the MRCP, and the MRCP(UK) is currently working on studies validating the examination, involving myself, Jane Dacre, Dr Katherine Woolf of DoME/ACME, and a PhD student in DoME/ACME, Kasia Ludka, supervised by Katherine and myself. I am based in both DoME/ACME and in Psychology.

The project in general involves looking at the relation of MRCP(UK) examination results to career outcomes of the doctors, as well as predictors of MRCP(UK) performance. In particular:

i) In relation to whether and if so when doctors go on to the Specialist Register of the General Medical Council's Register (now LRMP). The LRMP data are in the public domain.

ii) In relation to subsequent performance on the examinations of the Royal College of Radiologists, in particular Clinical Oncology, for which MRCP is a prior requisite. The RCR is collaborating in this research and has given permission for the data to be used in this way since the results would be of interest to them in evaluating their examinations.

iii) In relation to work-placed based assessments (WPBAs) collected by NHS Education Scotland (NES). NES have locally been informed that this usage of their data does not require ethical permission. NES is keen to assess its educational assessments in relation to other measures of educational achievement.

iv) Similar projects involving linkage to other outcome measures. All of the studies will involve linkage to existing databases collected for assessment or other educational purposes, and will be similar in format and content to those described above.

It should be emphasised that any publication of these data will only involve anonymised or aggregated data in which individuals cannot be identified, and all data will be kept securely and used only for the purposes of these studies. We will not be contacting NHS trusts or looking at the specific work carried out by these doctors in their employment by the NHS, but will only be considering educational measures collected as part of assessment or appraisal.

We would be grateful if you could let us know whether a formal application to the ethics committee is required.

With thanks,

Chris McManus  
Professor of Psychology and Medical Education



## Appendix B

### Systematic Review of the Literature

The review was performed based on a similar methodology to the one presented by Hutchinson *et al.* (2002). The key phrases were “predictive”, “validity”, “prediction”, “psychometric properties”, paired with “medical exams”, “medical assessments”, “medical”, “exams”. The medium of search was Google Scholar, which allows for the review of all major journal databases. The list of papers uncovered in the systematic review, with reference to the area of validity or psychometric quality that was covered in those papers, is presented below (ordered by the publication date). It is acknowledged that this publication list may not be exhaustive.

**Table B1. List of publications from a systematic review on psychometric properties of medical education related tests.**

<i>Paper</i>	<i>Aspect of validity</i>	<i>Exam referred</i>
1. Nibert, Young, & Adamson, 2002	Predictive validity	NCLEX, HESI
2. Tamblyn <i>et al.</i> , 2002	Predictive validity	QLEX
3. Norcini, Lipner, & Kimball, 2002	Predictive validity	ABIM certification data
4. Sharp, Bashook, Lipsky, Horowitz, & Miller, 2002	Predictive validity	Metanalysis ABIM
5. Simon, Volkan, Hamann, Duffey, & Fletcher, 2002	Predictive validity	OSCE, USMLE 1
6. Wragg, Wade, Fuller, Cowan, & Mills, 2003	Validity, reliability feasibility	miniCEX, DOPS, 360
7. Dacre, Besser, & White, 2003	Inter-rater consistency	PACES
8. Boulet <i>et al.</i> , 2003	Reliability validity	Simulation based acute care skills assesement
9. Ramos, Schafer, & Tracz, 2003	Reliability, validity	Fresno test of competence
10. McManus, Mooney-Somers, Dacre, & Vale, 2003	Reliability	Part 1 MRCP
11. Mcmanus, Smithers, Partridge, Keeling, & Fleming, 2003	General	A-levels and medical career
12. Norcini, Blank, Duffy, & Fortna, 2003	Assessment , reliability of ratings	miniCEX
13. Gonnella, Erdmann, & Hojat, 2004	Predictive validity	Medical school grades
14. Boenink, Oderwald, De Jonge, Van	Reliability, validity	Self-observe

<i>Paper</i>	<i>Aspect of validity</i>	<i>Exam referred</i>
Tilburg, & Smal, 2004		instrument
15. Downing, 2004	general	General
16. Julian, 2005	Predictive validity	MCAT
17. McManus, Mollon, Duke, & Vale, 2005	Item related analysis, cohort analysis	P 1 MRCP
18. McManus & Lissauer, 2005	Pass-Fail Decisions/ Cheating	RCPCH
19. Cohen, 2006	General	General
20. Grosch, 2006	Predictive validity	Metanalysis
21. Hojat <i>et al.</i> , 2007	General	General
22. West <i>et al.</i> , 2007	General,	Medical knowledge and professionalism
23. Norcini & Mckinley, 2007	General	General
24. Donnon, Paolucci, & Violato, 2007	Predictive	MCAT
25. Tamblyn <i>et al.</i> , 2007	Predictive validity	USMLE, CSE
26. Dewhurst, McManus, Mollon, Dacre, & Vale, 2007	Pass-rates / consequences	MRCP
27. Wilkinson <i>et al.</i> , 2008	Reliability feasibility	miniCEX, DOPS, MSF
28. Papadakis, Arnold, Blank, Holmboe, & Lipner, 2008	Predictive validity	Internal Medicine Residency Training
29. McManus <i>et al.</i> , 2008	Validity	MRCP
30. Gandy, Herial, Khuder, & Metting, 2008	Predictive validity	Academic performance
31. Walsh, Bailey, & Koren, 2009	Validity	OSCE
32. Pant, Nesargikar, & Cocker, 2009	Reliability, inter-rater consistency	TAB
33. Davies, Archer, Southgate, & Norcini, 2009	Generalisibility	F1 programme
34. Wenghofer <i>et al.</i> , 2009	Predictive validity	QE1, QE2
35. Al-Mahroos, 2009	Construct validity, generalisability	Clerkship evaluation in pediatrics
36. Levy, Mohanaruban, & Smith, 2010	Predictive validity	MRCP – WBA
37. Reid, Friedberg, Adams, McGlynn, & Mehrotra, 2010	Predictive validity	Certification, RAND quality

<i>Paper</i>	<i>Aspect of validity</i>	<i>Exam referred</i>
		assessment tools
38. Coumarbatch, Robinson, Thomas, & Bridge, 2010	Predictive validity	MCAT, academic performance
39. Tighe, McManus, Dewhurst, Chis, & Mucklow, 2010	General, reliability	MRCP P1 & 2
40. Hawkins, Margolis, Durning, & Norcini, 2010	Validity	MiniCEX
41. Zhao, Oppler, Dunleavy, & Kroopnick, 2010	Predictive validity, validity	MCAT
42. Puhan, Sinharay, Haberman, & Larkin, 2010	Internal structure/ General	Licensure exams
43. Elder <i>et al.</i> , 2011	Assessment, pass-fail rate	PACES
44. McManus, Ferguson, Wakeford, Powis, & James, 2011	Predictive validity	BMAT
45. Reinders <i>et al.</i> , 2011	Reliability	Consultation skills/ actors
46. Kobrin, Kim, & Sackett, 2011	Predictive validity	SAT
47. Prideaux <i>et al.</i> , 2011	Predictive validity	Selection procedures: MCAT, GAMSAT, UMAT, UKCAT
48. McGaghie, Cohen, & Wayne, 2011	Validity	USMLE Step 1 and 2
49. Mercer & Puddey, 2011	Predictive validity	Admissions criteria
50. Wilkinson, Zhang, & Parker, 2011	Predictive validity	Admissions test: UMAT
51. Speyer, Pilz, Van Der Kruis, & Brunings, 2011	Reliability, validity	Peer assessment of professional behaviours in medicine
52. Mitchell, Bhat, Herbert, & Baker, 2011	Predictive validity	WBAs
53. Lievens & Patterson, 2011	Validity, predictive validity	Situational judgment tests, knowledge tests, assessment centres
54. Ahmed <i>et al.</i> , 2012	Predictive validity	MRCGP (AKT and CSA)
55. Koczwara <i>et al.</i> , 2012	Validity	Situational Judgment Tests

<i>Paper</i>	<i>Aspect of validity</i>	<i>Exam referred</i>
		and clinical problem-solving tests
56. Hammond, O'Rourke, Kelly, Bennett, & O'Flynn, 2012	Validity, reliability	Dundee Ready Education Environment Measure
57. Yeung <i>et al.</i> , 2012	Reliability	FRCR2 oral exam
58. Metcalfe, 2012	Reliability, validity	MRCPGP AKT
59. Tiffin, Dowell, & McLachlan, 2012	Validity	UKCAT
60. Eva <i>et al.</i> , 2012	Predictive validity	Admissions procedures : Multiple mini-interviews
61. Poole, Shulruf, Rudland, & Wilkinson, 2012	Predictive validity	UMAT and GPA
62. Lievens & Sackett, 2012	Predictive validity	Situational Judgment Tests
63. Yoho, Antonopoulos, & Vardaxis, 2012	Predictive validity	GPA, MCAT, academic performance
64. McManus, Dewberry, <i>et al.</i> , 2013	Predictive validity	Student selection: UKCAT, A-levels, GCSEs, aptitude tests
65. McManus, Woolf, <i>et al.</i> , 2013	Predictive validity	GCSEs, A-levels, medical school performance, aptitude tests
66. Tiffin, McLachlan, Webster, & Nicholson, 2014	Predictive validity	PLAB
67. McManus & Wakeford, 2014	Predictive validity	PLAB
68. Glaros, Hanson, & Adkison, 2014	Predictive validity	Academic performance
69. Pugh <i>et al.</i> , 2014	Validity, reliability	OSCE

## Appendix C

### Example PACES scenario

#### INFORMATION FOR THE CANDIDATE

Scenario N° EX1

#### MRCP(UK) PACES

##### Station 2: HISTORY TAKING

<b>Patient details:</b>	Mrs Heba Kamel, a 54-year-old woman
<b>Your role:</b>	The doctor in the general medical outpatient clinic
<b>Presenting complaint:</b>	Progressively worsening dyspnoea

Please read the letter printed below. When the bell sounds, enter the room. You have 14 minutes to take a history from the patient, 1 minute to collect your thoughts and 5 minutes for discussion. You may make notes if you wish.

##### Referral text:

Dear Doctor,

This retired nurse has had progressively worsening dyspnoea for the past 18 months. She has a history of recurrent urinary tract infections and is on long-term antibiotic therapy.

She smokes 20 cigarettes per day and has done so for the past 20 years. She has no past respiratory history.

She has hypertension and is known to have right bundle branch block on her ECG. Examination reveals definite bi-basal crackles on auscultation of the chest. Full blood count and urea and electrolytes are normal.

Please see and advise on her management.

Your sincerely,

**Your task** is to interview the patient and, based on the history you obtain, construct a differential diagnosis and plan for investigation. You should explain these to the patient and answer any questions they may have.

##### DO NOT EXAMINE THE PATIENT

Any notes you make must be handed to the examiners at the end of the station



## Appendix D

### Additional Tables and Graphs to Chapter 4

#### SECTION 4.1

**Table D1. Frequencies of passed and failed MRCP(UK) Part I candidates based on the number of attempts.**

		<i>Number of attempts</i>								<i>Subtotal</i>
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8 to 26</i>	
failed	N	8,386	2,955	1,386	7,93	455	260	153	361	14,749
	%	21.3%	7.5%	3.5%	2.0%	1.2%	0.7%	0.4%	0.9%	37.5%
passed	N	14,217	4,875	2,398	1,305	769	407	242	373	24,586
	%	36.1%	12.4%	6.1%	3.3%	2.0%	1.0%	0.6%	0.8%	62.5%
Subtotal	N	22,603	7,830	3,784	2,098	1,224	667	395	734	39,335
	%	57.5%	19.9%	9.6%	5.3%	3.1%	1.7%	1.0%	1.7%	100.00%

% are calculated from the total number of valid cases.

**Table D2. Frequencies of passed and failed MRCP(UK) Part II candidates based on the number of attempts.**

		<i>Number of attempts</i>								<i>Subtotal</i>
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8 to 21</i>	
failed	N	1,195	511	285	136	89	56	33	72	2,377
	%	5.06%	2.16%	1.21%	0.58%	0.38%	0.24%	0.14%	0.30%	10.06%
passed	N	14,512	3,648	1,505	717	370	197	130	181	21,260
	%	61.40%	15.43%	6.37%	3.03%	1.57%	0.83%	0.55%	0.77%	89.94%
Subtotal	N	15,707	4,159	1,790	853	459	253	163	253	23,637
	%	66.45%	17.60%	7.57%	3.61%	1.94%	1.07%	0.69%	1.07%	100.00%

\*% are calculated from the total number of valid cases.

**Table D3. Frequencies of passed and failed MRCP(UK) PACES candidates based on the number of attempts.**

		Number of attempts								Subtotal
		1	2	3	4	5	6	7	8 to 14	
failed	N	1,326	812	786	244	166	162	62	83	3,641
	%	6.23%	3.82%	3.70%	1.15%	0.78%	0.76%	0.29%	0.39%	17.12%
passed	N	10,353	4,014	1,788	719	389	214	81	71	17,629
	%	48.67%	18.87%	8.41%	3.38%	1.83%	1.01%	0.38%	0.33%	82.88%
Subtotal	N	11,679	4,826	2,574	963	555	376	143	154	21,270
	%	54.91%	22.69%	12.10%	4.53%	2.61%	1.77%	0.67%	0.72%	100.00%

\*% are calculated from the total number of valid cases.

## SECTION 4.3

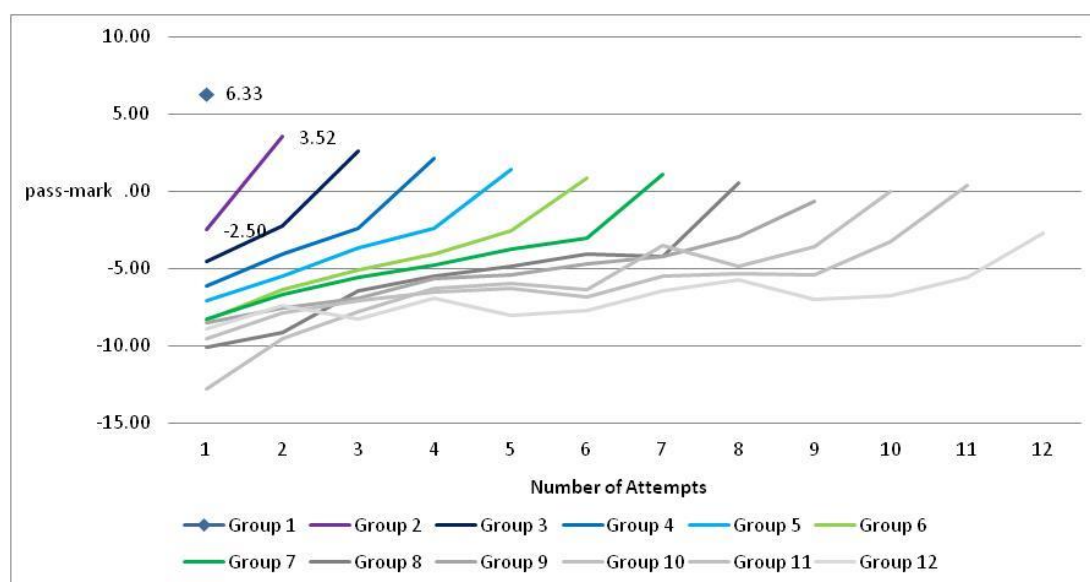


Figure D1. Mean scores in Part II per attempt for groups based on total number of attempts – approximation of the learning curve.



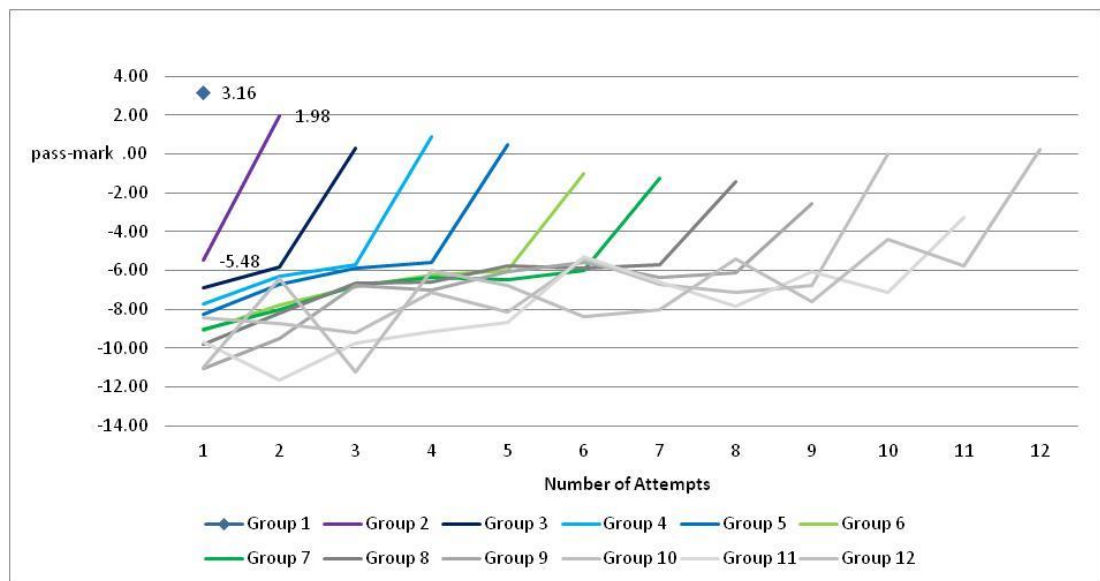


Figure D2. Mean scores in PACES per attempt for groups based on total number of attempts – approximation of the learning curve

**Table D4. Mean scores for the twelve groups based on the Total Number of Attempts in Part I (showing the learning process).**

<i>No of attempt</i>	<i>Group 1</i>	<i>Group 2</i>	<i>Group 3</i>	<i>Group 4</i>	<i>Group 5</i>	<i>Group 6</i>	<i>Group 7</i>	<i>Group 8</i>	<i>Group 9</i>	<i>Group 10</i>	<i>Group 11</i>	<i>Group 12</i>
Attempt 1	-0.1	-8.5	-11.2	-13.0	-14.9	-16.0	-17.4	-18.2	-18.3	-19.2	-19.0	-17.8
Attempt 2		-1.5	-7.4	-9.1	-11.4	-12.4	-13.5	-15.2	-16.8	-16.2	-17.4	-14.8
Attempt 3			-1.4	-6.7	-8.5	-9.8	-11.8	-12.9	-13.8	-15.4	-15.8	-13.4
Attempt 4				-1.4	-6.6	-7.8	-9.5	-11.7	-11.9	-13.5	-13.0	-12.1
Attempt 5					-1.6	-6.0	-7.8	-9.2	-10.8	-11.4	-11.4	-12.8
Attempt 6						-1.7	-6.2	-7.4	-9.4	-9.5	-11.8	-12.2
Attempt 7							-1.7	-7.0	-8.6	-9.0	-10.7	-9.4
Attempt 8								-2.6	-7.8	-8.0	-8.9	-9.9
Attempt 9									-4.3	-7.0	-7.3	-8.3
Attempt 10										-4.1	-6.4	-7.1
Attempt 11											-3.3	-6.4
Attempt 12												-3.5

**Table D5. Mean scores for the twelve groups based on the Total Number of Attempts in Part II (showing the learning process).**

<i>No of Attempts</i>	<i>Group 1</i>	<i>Group 2</i>	<i>Group 3</i>	<i>Group 4</i>	<i>Group 5</i>	<i>Group 6</i>	<i>Group 7</i>	<i>Group 8</i>	<i>Group 9</i>	<i>Group 10</i>	<i>Group 11</i>	<i>Group 12</i>
Attempt 1	6.33	-2.50	-4.58	-6.14	-7.08	-8.36	-8.31	-10.08	-8.49	-12.77	-9.52	-8.88
Attempt 2		3.52	-2.29	-4.07	-5.53	-6.41	-6.73	-9.15	-7.59	-9.58	-7.87	-7.38
Attempt 3			2.55	-2.38	-3.65	-5.07	-5.57	-6.44	-6.89	-7.80	-7.09	-8.25
Attempt 4				2.12	-2.39	-4.08	-4.82	-5.51	-5.67	-6.30	-6.57	-6.94
Attempt 5					1.42	-2.55	-3.75	-4.86	-5.41	-6.01	-6.29	-8.04
Attempt 6						0.80	-3.01	-4.10	-4.68	-6.35	-6.87	-7.69
Attempt 7							1.09	-4.26	-4.27	-3.53	-5.53	-6.49
Attempt 8								0.54	-2.98	-4.88	-5.35	-5.70
Attempt 9									-0.63	-3.62	-5.39	-7.00
Attempt 10										-.05	-3.28	-6.75
Attempt 11											0.34	-5.54
Attempt 12												-2.70

**Table D6. Mean scores for the twelve groups based on the Total Number of Attempts in PACES (showing the learning process).**

<i>No of attempts</i>	<i>Group 1</i>	<i>Group 2</i>	<i>Group 3</i>	<i>Group 4</i>	<i>Group 5</i>	<i>Group 6</i>	<i>Group 7</i>	<i>Group 8</i>	<i>Group 9</i>	<i>Group 10</i>	<i>Group 11</i>	<i>Group 12</i>
Attempt 1	3.16	-5.48	-6.89	-7.74	-8.26	-9.07	-9.03	-9.80	-11.06	-8.43	-9.67	-11.00
Attempt 2		1.98	-5.85	-6.32	-6.70	-7.80	-8.04	-8.21	-9.48	-8.71	-11.67	-6.40
Attempt 3			0.28	-5.73	-5.90	-6.85	-6.76	-6.63	-6.79	-9.21	-9.73	-11.20
Attempt 4				0.91	-5.60	-6.22	-6.36	-6.57	-7.00	-7.14	-9.13	-6.00
Attempt 5					0.50	-6.00	-6.50	-5.79	-6.04	-8.14	-8.67	-6.80
Attempt 6						-1.00	-6.01	-5.88	-5.59	-5.43	-5.27	-8.40
Attempt 7							-1.26	-5.68	-6.35	-6.70	-6.60	-8.00
Attempt 8								-1.42	-6.11	-7.12	-7.87	-5.40
Attempt 9									-2.54	-6.78	-6.07	-7.60
Attempt 10										-0.02	-7.14	-4.38
Attempt 11											-3.30	-5.78
Attempt 12												0.24

**Table D7. Results of the post-hoc REGW Q test for one-way ANOVA on first attempt Part II scores with the twelve groups based on the Total Number of Attempts being the factor.**

Attempts / Group No.	N	Homogeneous groups				
		1 ( <i>p</i> =0.437)	2 ( <i>p</i> =0.099)	3 ( <i>p</i> =1.0)	4 ( <i>p</i> =1.0)	5 ( <i>p</i> =1.0)
10	10	-12.77				
12	12	-10.25				
8	38	-10.07				
11	12	-9.52				
9	32	-8.49				
6	170	-8.36				
7	100	-8.31	-8.31			
5	330	-7.08	-7.078			
4	653		-6.14			
3	1,529			-4.58		
2	3,805				-2.50	
1	15,707					6.33
Levene's test: $F(11, 22,386) = 65.00, p < 0.001$						
One-Way ANOVA: : $F(11, 22,386) = 1,336.72, p < 0.001$						

**Table D8. Results of the post-hoc REGW Q test for one-way ANOVA on first attempt Part II scores with the twelve groups based on the Total Number of Attempts being the factor.**

Attempts / Group No.	N	Homogeneous groups				
		1 ( <i>p</i> =0.142)	2 ( <i>p</i> =0.310)	3 ( <i>p</i> =1.0)	4 ( <i>p</i> =1.0)	5 ( <i>p</i> =1.0)
9	48	-11.06				
12	7	-11.00				
8	70	-9.80				
11	15	-9.67				
6	376	-9.07				
7	143	-9.03	-9.03			
10	14	-8.43	-8.43			
5	555	-8.26	-8.26			
4	961		-7.74			
3	2,571			-6.89		
2	4,821				-5.48	
1	11,679					3.16
Levene's test: $F(11, 21,248) = 7.65, p < 0.001$						
One-Way ANOVA: : $F(11, 21,248) = 1,903.75, p < 0.001$						

## Appendix E

### Meta-analyses: Funnel Plots

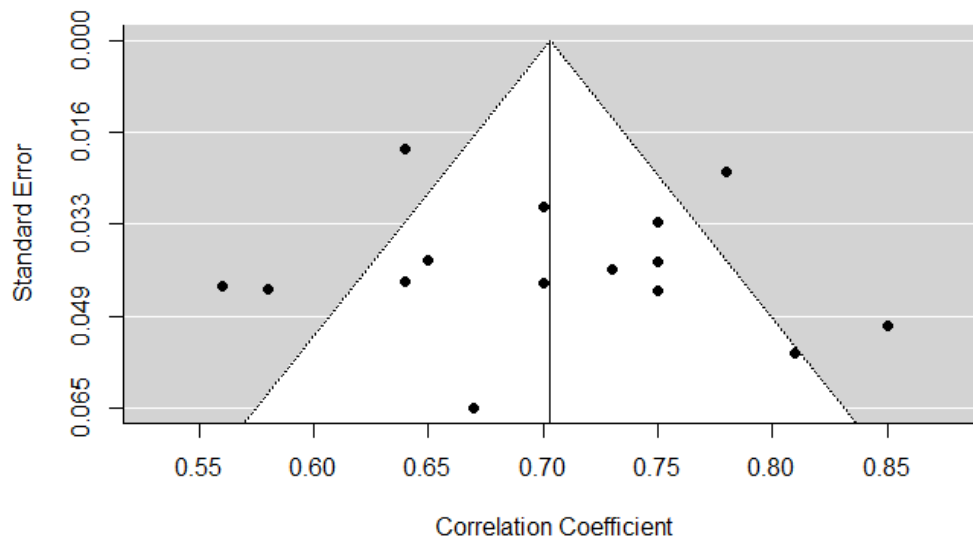


Figure E1. Funnel plot for meta-analytical model on coefficients associated with Part II and examinations.

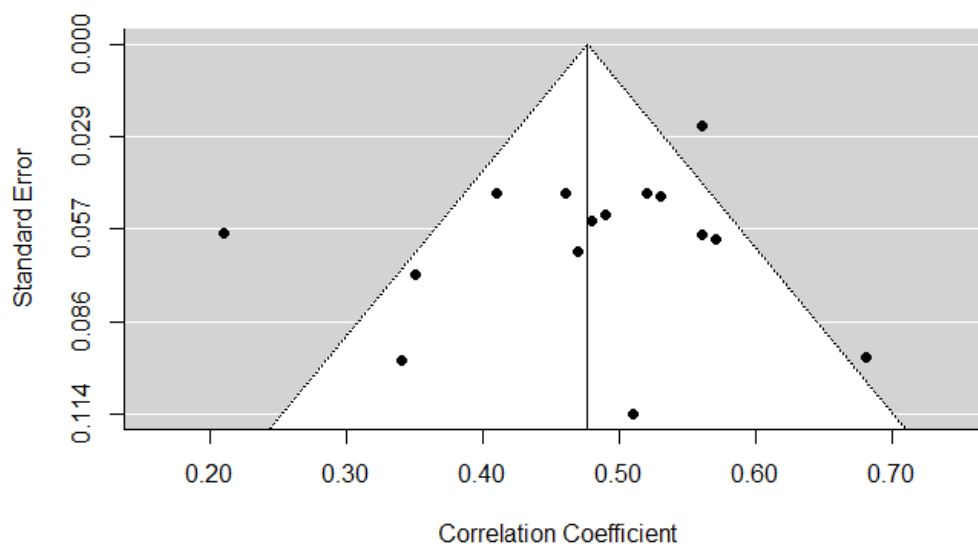


Figure E2. Funnel plot for meta-analytical model on coefficients associated with PACES and examinations.

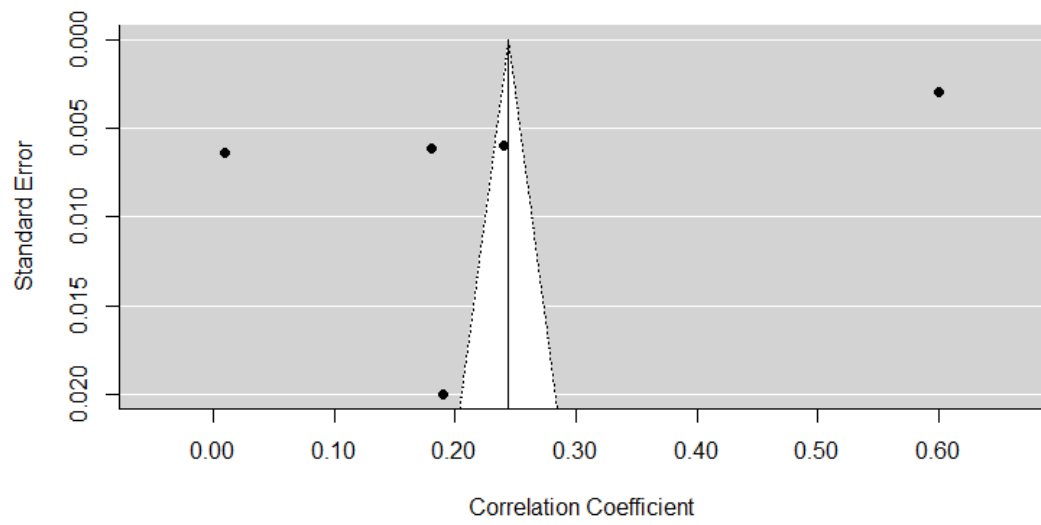


Figure E3. Funnel plot for meta-analytical model on coefficients associated with Part I and underperformance criteria.

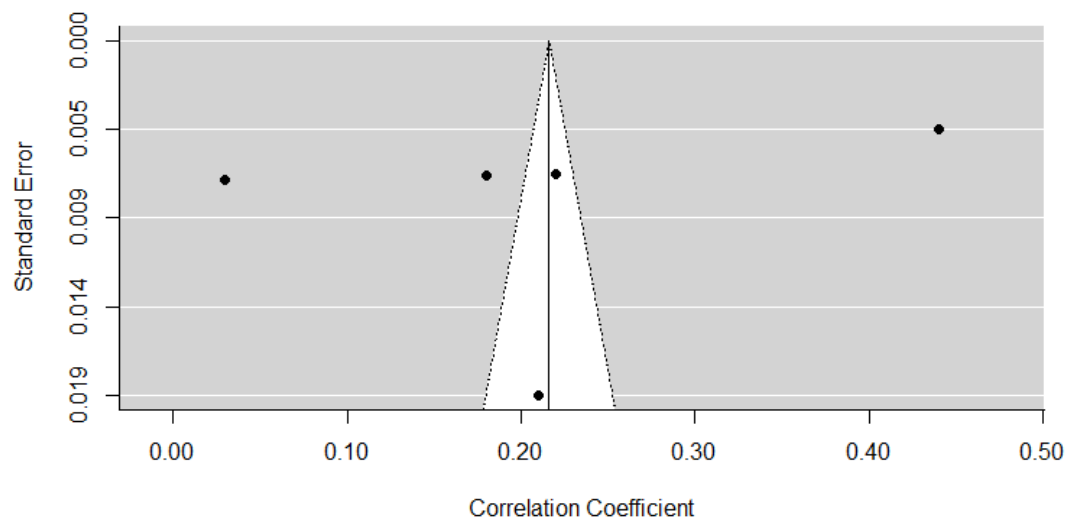


Figure E4. Funnel plot for meta-analytical model on coefficients associated with Part II and underperformance criteria.

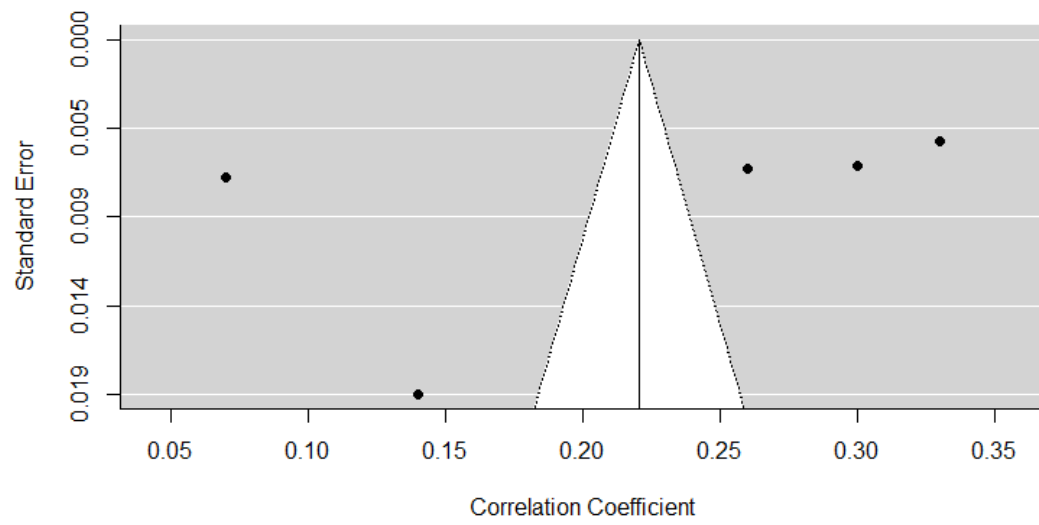


Figure E5. Funnel plot for meta-analytical model on coefficients associated with PACES and underperformance criteria.